Geodata and Analysis:

Statistics of one variable

data quality

- precision and accuracy, box1
- data entry
  - missing data - * or absurd value
  - labels, code - E&N coordinate
  - format - text, spreadsheet, ASCII
    - (create ascII file in PC), import to excel

types of geodata

- **ratio** - measures of length and weight etc
- **interval** - can be negative (C,F,K)
- **ordinal** - intervals not regular e.g. Moh's hardness and. Richter's earthquake scales
- **directional** - angles

Types of analysis

- **deterministic** - output determined
- **stochastic** - output uncertain
  - descriptive statistics
  - univariate (one variable)
  - bivariate
  - multivariate analysis
  - **time series**
  - **spatial analysis**

Sampling strategies

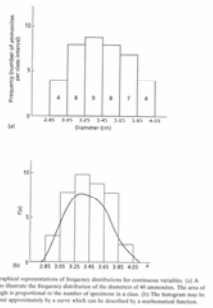
- sampling size
- spatial schemes
  - regular grided (rec and tri)
  - ununiform (rnd within grid)
  - random
  - clustered
  - traverse

Creation of Histogram

- pdf – Probability Density Function or histogram
- Source data (time or space series)
- Determine the sizes of Bins to store data of fixed ranges
- Sorting (a-z or z-a)
- Count the # of data fitting each Bin (i)
- Probability Pi=#i/(1+total # of data)
- Plot Pi – Bin size

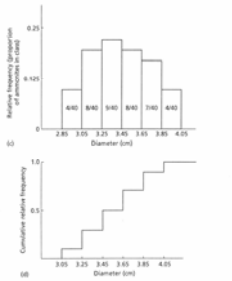
Descriptive Statistics

- **Frequency distribution**
  - pdf – frequency density function f(x)
- **Histogram** – a graphics expression (7 steps)
  1. Determine the number of bins (square root of the total number or use natural boundaries)
  2. Calculate the bin size =(max-min)/ (# of bins), or use natural boundaries
  3. Counts in each Bin
  4. Covert counts to % height or density (counts/total counts)
  5. Draw axes and label
  6. Draw in bars
  7. Draw the cumulative frequency



Descriptive Statistics

- **Relative frequency** (determined by the size of the population)
  - Total area under the histogram is 1
  - Cumulative frequency – maximum value is 1.
- **Cumulative frequency distribution**
  - Not affected by the bin width.
  - Do not show pdf so clearly



Properties of Frequency Distribution

- **Location** – average value
- **Dispersion** – extent of variation
- **Shape** – symmetry

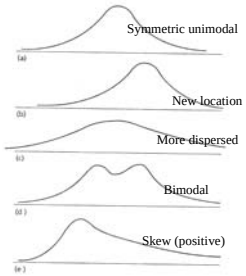


Figure 2.3

Parameters of Population Frequency - Position

- **Arithmetic mean**:

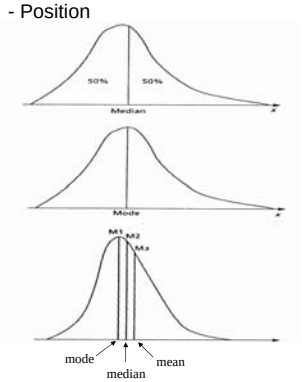
$$m = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N}$$

- **Geometric mean** (for non-symmetric or skew populations): =GEOMEAN(A1..A100)

$$g = \left(\prod(X_i)\right)^{\frac{1}{N}} = \left(X_1 \cdot X_2 \cdot X_3 \cdots X_N\right)^{\frac{1}{N}}$$

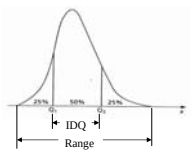
Parameters of Population Frequency - Position

- Median (for non-symmetric or skew populations) – middle value on the sorted list.
- Mode – most frequently occurring value in a population, at which the height of pdf is the greatest



Parameters of Population Frequency - dispersion

- **Range** = Xmax - Xmin
- **IDQ**: Interquartile Deviation
  IDQ=X(Q_3)-X(Q_1)
- **Variance and Standard Deviation**



mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N}$

Variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 = \frac{1}{N}\left[(X_1 - \mu)^2 + (X_2 - \mu)^2 + \ldots (X_N - \mu)^2\right]$

Standard deviation (SD): $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$

Root of mean square (RMS) of (deviation from average)

$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \mu^2}$

Square Root of (average of squares – square of average)

Parameters of Population Frequency
- dispersion

Standard deviation (SD) ***2nd formula (quicker and easier)***:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^2} = \sqrt{\frac{\sum_{i=1}^{N}\left(X_i^2 - 2\mu X_i + \mu^2\right)}{N}} = \sqrt{\frac{\sum_{i=1}^{N}X_i^2 - \sum_{i=1}^{N}(2\mu X_i) + \sum_{i=1}^{N}\mu^2}{N}}$$

$$= \sqrt{\frac{\sum_{i=1}^{N}X_i^2}{N} - 2\mu\frac{\sum_{i=1}^{N}X_i}{N} + \frac{N\mu^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}X_i^2}{N} - 2\mu\mu + \mu^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \mu^2}$$

$$\mu = \frac{\sum_{i=1}^{N}X_i}{N} \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \mu^2}$$

Square Root of (average of squares – square of average)

Parameters of Population Frequency
- dispersion

Using the ***2nd formula of SD (quicker and easier)***: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \mu^2}$

Exercise 1: Find SD of 1,3, 4, 5,7 (μ =4)

Exercise 2: Rock samples: Group 1 has 12 samples with the average diameter of 20 cm and SD of 3 cm;
Group 1 has 18 samples with the average diameter of 23 cm and SD of 4 cm.
Find the average size and SD of all 30 samples.

Parameters of Population Frequency
- shape

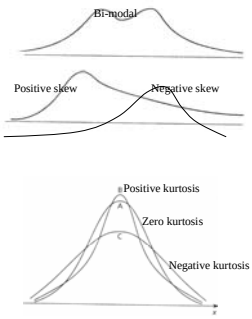
- **Number of modes**
  - Multi- or Poly-modal?
- **Skewness**

$$\text{Pearson skewness} = \frac{\text{Mean - Mode}}{\text{standard deviation}}$$

$$\text{Fisher skewness}: \gamma_{\text{skewness}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^3}{\sigma^3}$$

- **Kurtosis** - peakedness

$$\text{Coefficient of kurtosis}: \gamma_{\text{kurtosis}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i-\mu)^4}{\sigma^4} - 3$$



Parameters of Sample Frequency
- shape

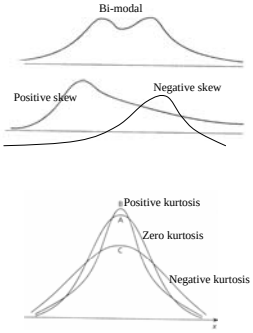
- **Number of modes**
  - Multi- or Poly-modal?
- **Skewness**

$$\text{Pearson skewness} = \frac{\text{Mean - Mode}}{\text{standard deviation}}$$

$$\text{Fisher skewness}: \gamma_{\text{skewness}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^3}{s^3}$$

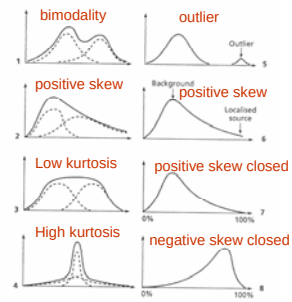
- **Kurtosis** - peakedness

$$\text{Coefficient of kurtosis}: \gamma_{\text{kurtosis}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^4}{s^4} - 3$$



Descriptive statistics of sample data - shape

- **Normal distribution** - unimodal, symmetric
- **Mixtures** - of more than one Normal distribution, need to decompose.
- **Outliers** - anomalous values (errors or unknown mechanisms)
- **Data closure** - closed data (%, ppm) reflect many artifacts causing skewed distributions. use log ratio transformation
- **Exponents** - power function exponents may cause a skew



Graphical data analysis

Why Graphics?

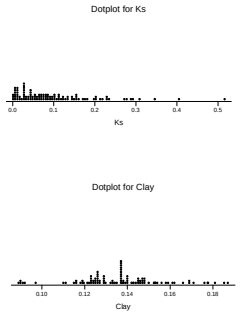
- Detect wrong inputs or values
- Type of distribution
- Need for sophisticated analysis
- Suggestion of hypotheses or models
- ...

Types of graphics

- Dot plot
- Stem-and-leaf diagram
- Box plots
- Scatter diagrams
- Marginal plot
- Matrix plot

Dot plot

- Similar to histogram
- Repeated values represented by dots on top of each other.



Stem-and-leaf diagram

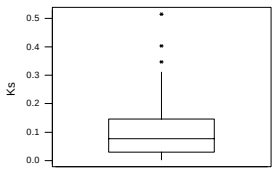
The individual values are divided into the most significant part (**Stem**) and the rest (**leaf**).

Stem-and-leaf of Ks N = 100
Leaf Unit = 0.010

```
 36    0 000000000001111112222222223334444444
(26)   0 55555666666777778888889999
 38    1 00011112223344
 24    1 555566779
 15    2 0012233
  8    2 7789
  4    3 14
  2    3
  2    4 0
  1    4
  1    5 1
```

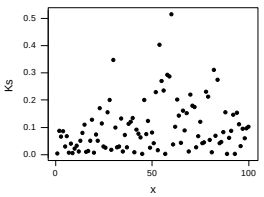
Box plots

- Indicating suspicious values and for comparing sets of data.
- Median is the line inside box.
- Lower hinge - midway between median and MIN
- Higher hinge - midway between median and max.
- Vertical line to extreme values (outlier or interesting ones).



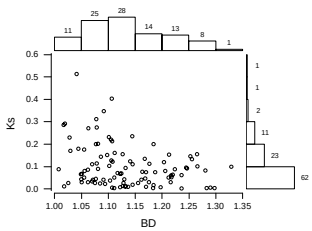
Scatter diagrams

- Trend and relationship, degree of correlation

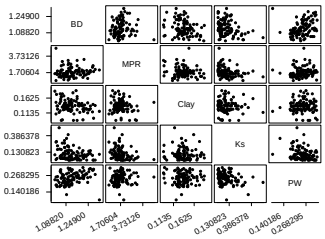


Marginal Plot

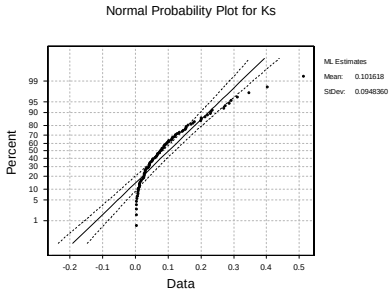
- Trend and relationship, histogram



Matrix Plot



Normal Probability Plot



Do you know what is normal probability? What is x and what is y in this plot?

Lab 1. Descriptive statistics

- Select a parameter data array from dataset
- Calculated **relative frequency, cumulative frequency, arithmetic and geometric means**.
- Determine **median, mode, range, IDQ, STD, number of modes, skewness and kurtosis**.
- Make various graphics.
- Print out the results on a PDF file with your name and lab number in the first page
- Upload the file to blackboard.