Groundwater Data and Statistical Analysis

Sampling and Data Collection

- **Groundwater data and chance!**
- **Data Quality**

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

- **Sources of uncertainty (is the data set good enough)?**
  - Natural randomness effect
    - topography,
    - lithological changes in the aquifer,
    - water quality changes in an aquifer … etc.

    ➔ may need more samples
  - Sampling errors
    - failure in measurement procedures,
    - small sampling size for a large potential problem … etc.

    ➔ follow proper guidelines
  - Incorrect understanding of hydrogeology and contamination process

When sampling, put into consideration:

- a) Number of samples

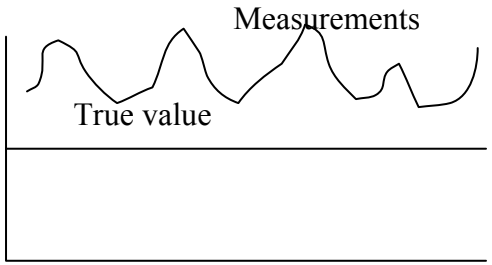
  ➔ higher # of samples = good results, more accuracy & precision (cost?)

- b) Global picture of the problem

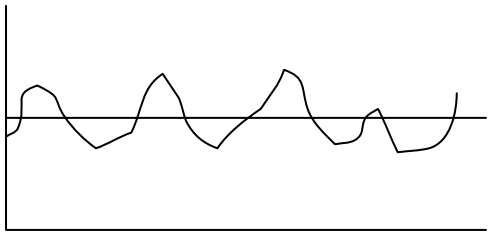
  ➔(i.e. understand geology, hydrogeology, economy …etc.)

- c) Calibration of measuring devices

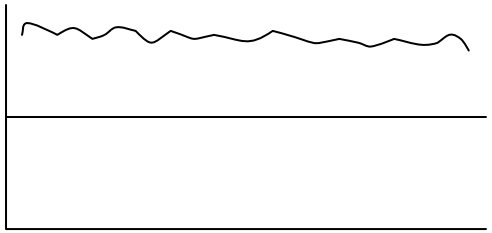
  ➔ precision and accuracy



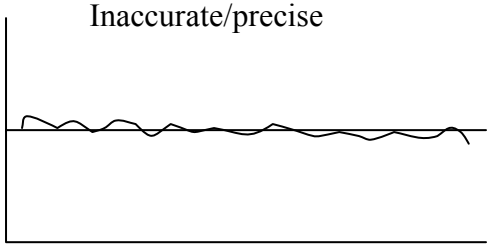
Inaccurate/non-precise



Accurate/non-precise



Inaccurate/precise



Accurate/precise

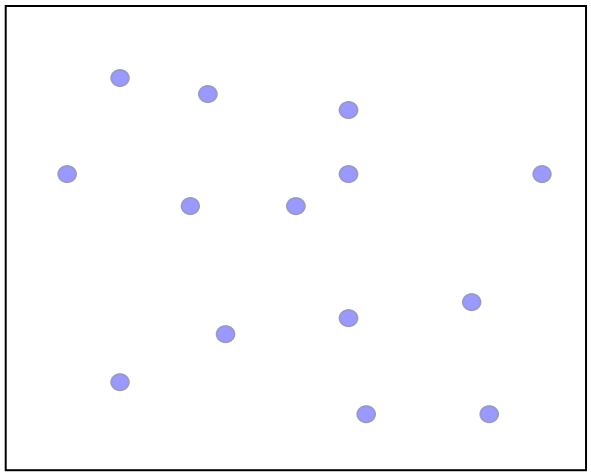
Sampling Strategies

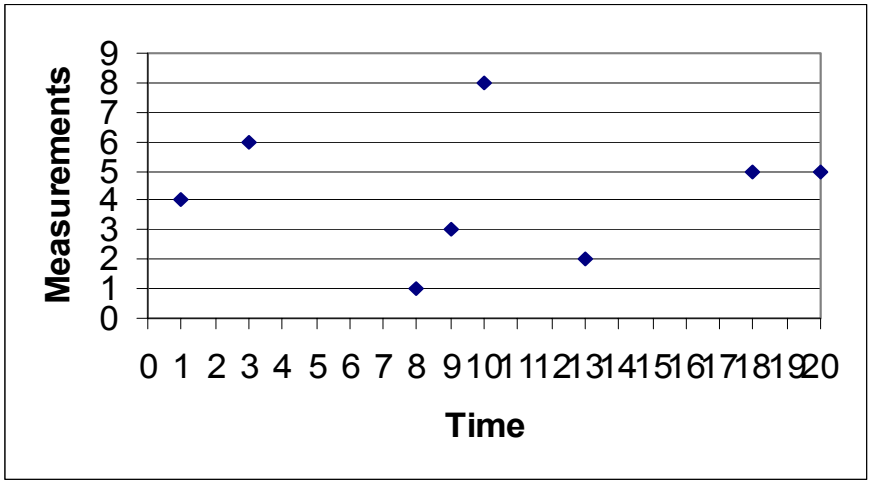
Sample size: Could be decided on the basis of a pilot study.

Spatial sampling schemes:

- **Random sampling**

  ➔ use random number generator to decide about sampling plan (e.g. where to place monitoring wells in a regional aquifer)

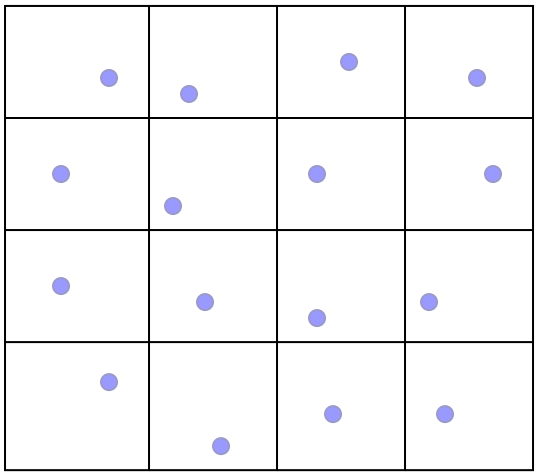


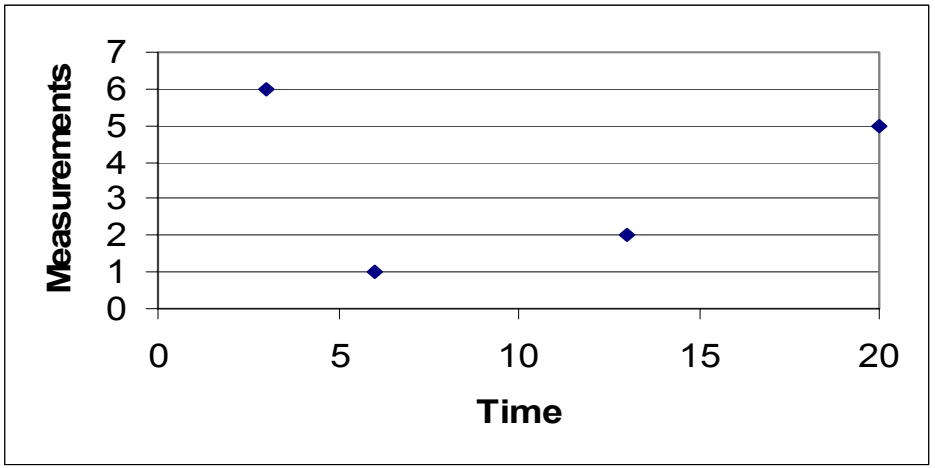


Stratified random sampling

➔ use randomization sampling process within pre defined spatial squares or times

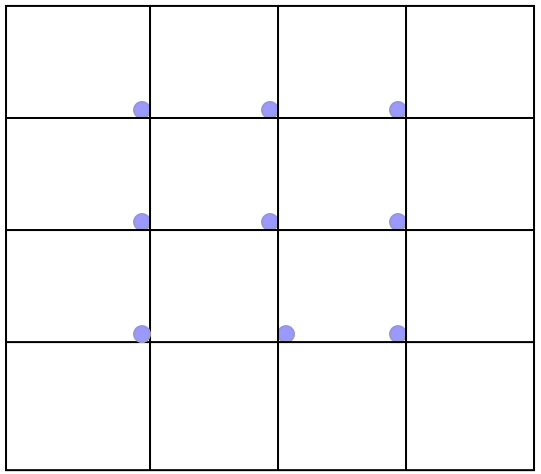
➔ Best to reduce bias

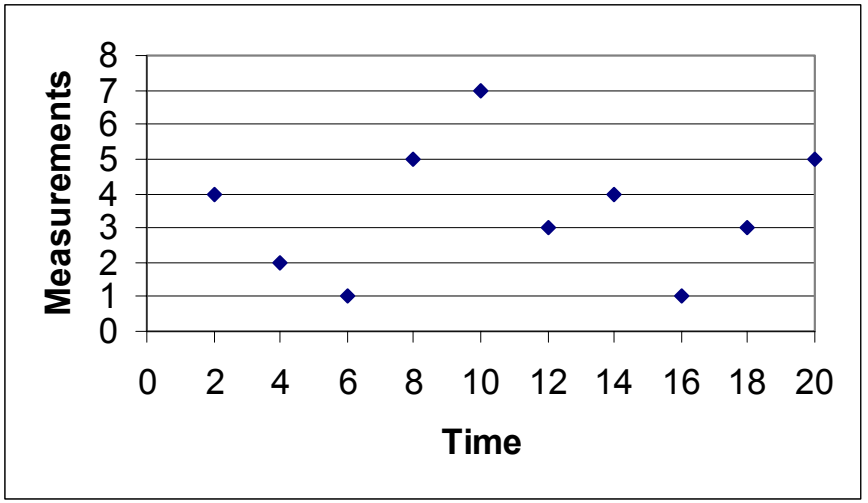




- **Regular (uniform) sampling**

  ➔ use regular spatial squares or rectangles or equally pre defined times to sample

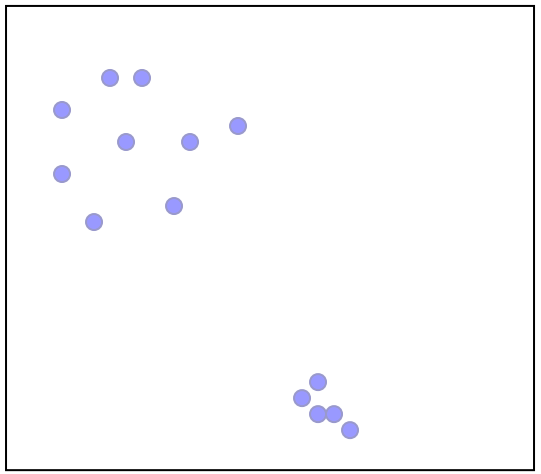


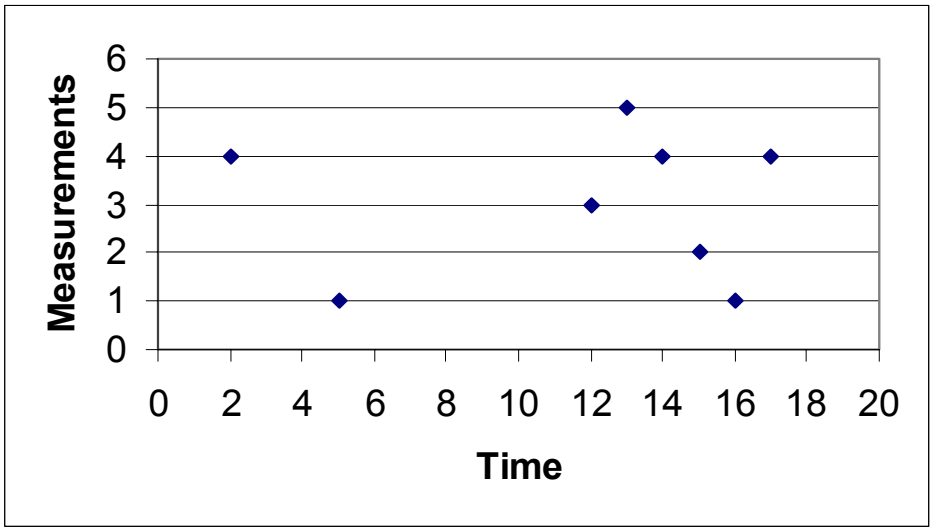


Clustered sampling

➔ usually used in sampling due to

- concentration of a contaminant in part of the aquifer or during specific period of time
- accessibility restrictions.





Populations and Samples

When making hydrogeological measurements and analysis, in what are we interested?

Are we interested in:

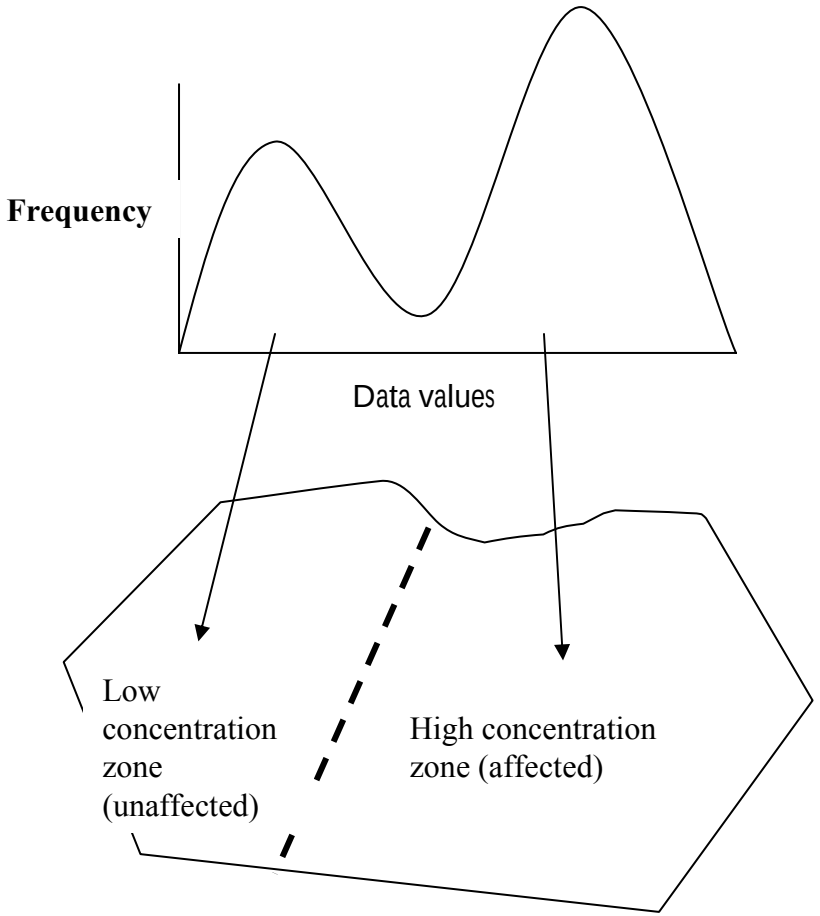
1. Properties of pieces of rocks, core samples, chemical concentrations … etc. **only** or
2. Properties of the hydrogeologic phenomenon under study (e.g. aquifer, plume … etc.

In fact, we are interested in understanding #2 or properties of the hydrogeologic **POPULATION**.

However, #1 is important because it represents a **SAMPLE** which is a subset of the **POPULATION**

Why Statistics?

1) Quantify hydro geological information from a data set (data Sample) to represent the Population (e.g. aquifer, plume … etc.)



Why Statistics?

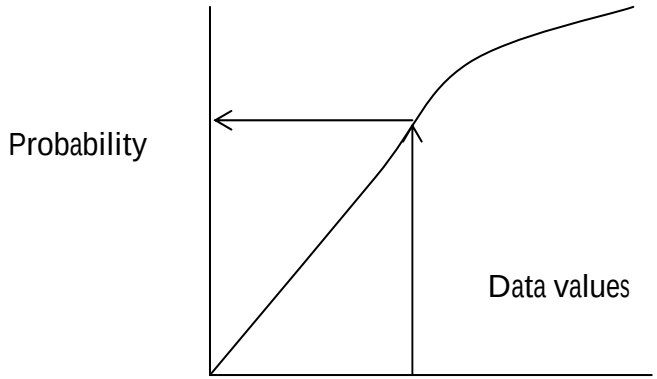
2) Conduct **feasibility studies**

➔porosity distribution in a geologic unit … etc.

3) Perform **uncertainty analysis & risk assessment**

➔ environmental site investigation, probability of existence of a contaminant in an aquifer … etc.

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.

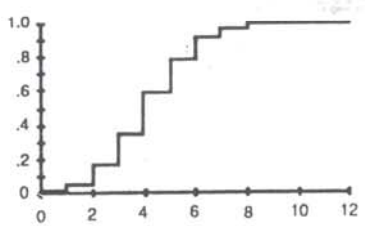
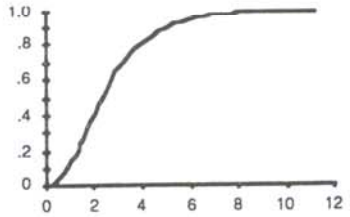
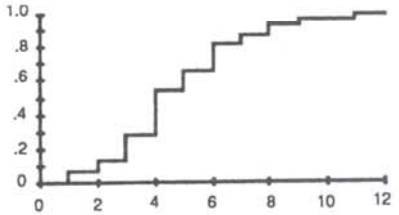
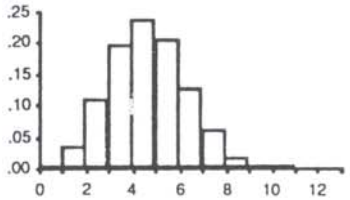
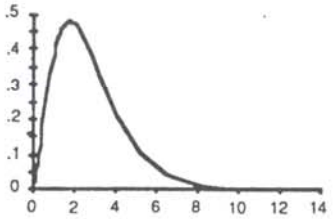
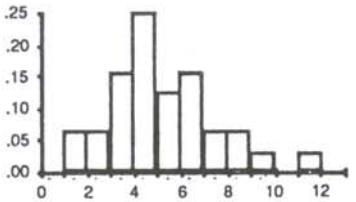
- Univariate statistics are represented by:
  - Summary statistics
  - Graphical plots

Summary Statistics

Statistics used to describe a sample (data set) correspond to parameters that are used to describe a population.

Summary Statistics

TABLE 17.2.1

Concept	Population value, discrete case	Population value, continuous case	Sample value
Cumulative distribution function (cdf)	 Describes the probability that a random variable is less than or equal to a specified value x	 Describes the probability that a random variable is less than or equal to a specified value x	 Empirical distribution function (edf): describes the observed frequency of a random variable being less than or equal to a specified value x
Probability mass function (pmf) and probability density function (pdf)	 pmf: the probability that X is equal to k	 pdf: first derivative of the cumulative distribution function	 Histogram: observed frequency with which random variable X falls into the assigned ranges
Mean, average, or expected value	$\mu \equiv \sum_{i=1}^{\infty} P(X = x_i)x_i$	$f(x) \equiv \frac{dF(x)}{dx}$ $\mu \equiv \int_{-\infty}^{\infty} xf(x)\,dx$	$\bar{X} \equiv \sum_{i=1}^{n} \frac{X_i}{n}$

Maidment, D., 1993, Handbook of Hydrology

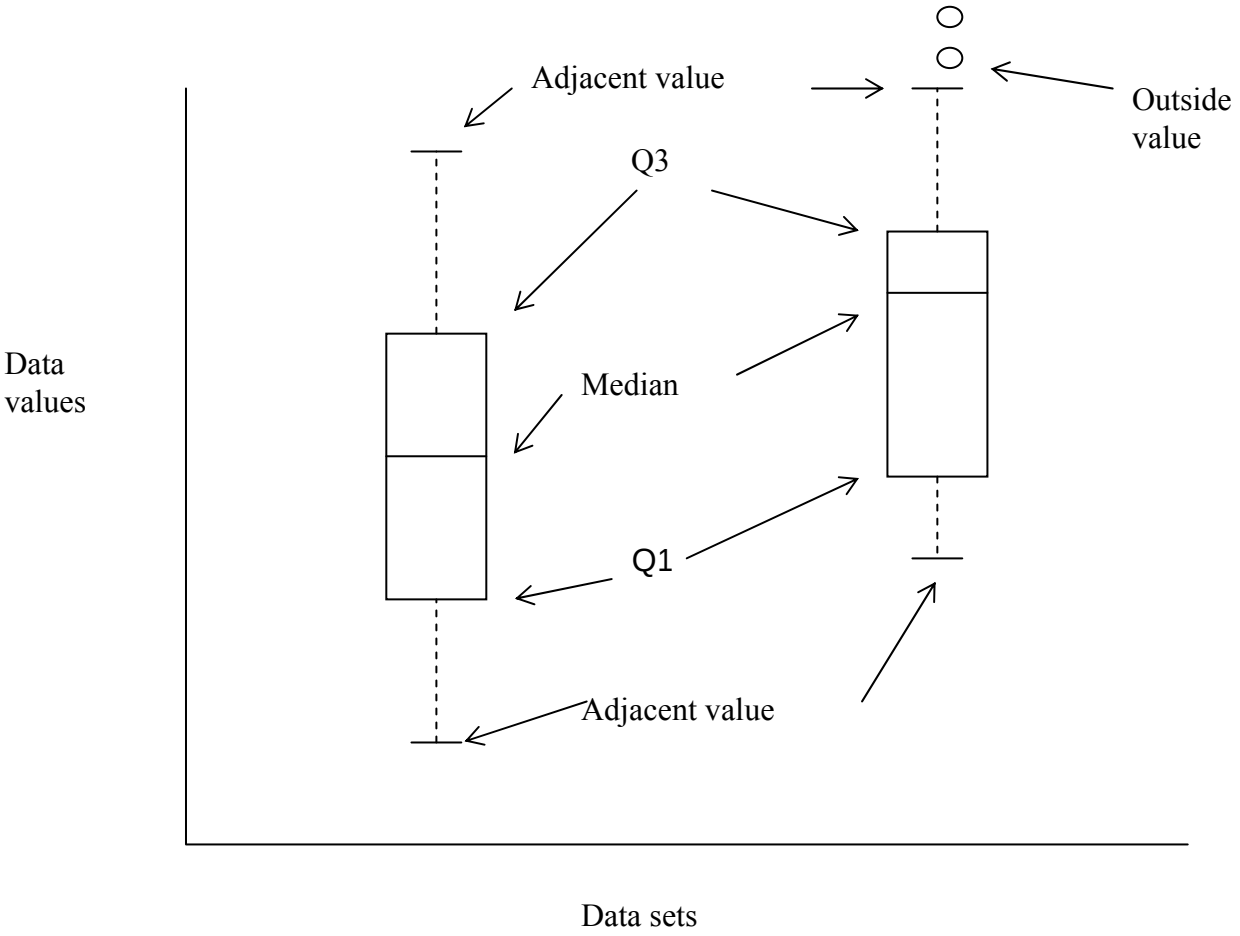
Summary Statistics

Variance	$\sigma^2 \equiv \sum_{i=1}^{\infty} P(X = x_i)(x_i - \mu)^2$	$\sigma^2 \equiv \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$	$S^2 \equiv \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$
kth central moment	$M_k \equiv \sum_{i=1}^{\infty} P(X = x_i)(x_i - \mu)^k$	$M_k \equiv \int_{-\infty}^{\infty} (x - \mu)^k f(x)\, dx$	$\tilde{M}_k \equiv \sum_{i=1}^{n} \frac{(X_i - \bar{X})^k}{n}$
Standard deviation	$\sigma \equiv \sqrt{\sigma^2}$		$S \equiv \sqrt{S^2}$
Coefficient of variation or relative standard deviation (if $\mu \neq 0$)	$\text{CV} \equiv \frac{\sigma}{\mu}$		$\text{CV} \equiv \frac{S}{\bar{X}}$
Coefficient of skew (a measure of asymmetry)	$\gamma \equiv \frac{M_3}{\sigma^3}$		$G \equiv \frac{\tilde{M}_3}{S^3}$
Quantiles	x_p is any value of X that has the properties that $P(X < x_p) \leq p$, $P[X > x_p] \leq 1 - p$		$\hat{X}_p$ is the pth quantile of EDF
Median (useful for describing central tendency regardless of skewness)	$x_{0.5}$ Any value of X that has the property that $P[X < x_p] \leq 0.5$, $P[X > x_p] \leq 0.5$		$\hat{X}_{0.5}$ The middle observation in a sorted sample, or the average of the two middle observations if the sample size is even.
Upper quartile, lower quartile, and hinges	Upper quartile $\equiv x_{0.75}$ Lower quartile $\equiv x_{0.25}$		Upper hinge $\equiv \hat{X}_{0.75}$ This is an approximation to the sample upper quartile; it is defined as the median of all sample values of $X \geq x_{0.50}$. The lower hinge, $\hat{X}_{0.25}$, is defined analogously.
Interquartile range (useful for describing spread of data regardless of symmetry)	$x_{0.75} - x_{0.25}$ Width of central region of population containing probability of 0.5		$\hat{X}_{0.75} - \hat{X}_{0.25}$ Width of central region of data set encompassing approximately half the data

Maidment, D., 1993, Handbook of Hydrology

Graphical Display of Data

Box - Whisker Plot



Graphical Display of Data

Box - Whisker Plot

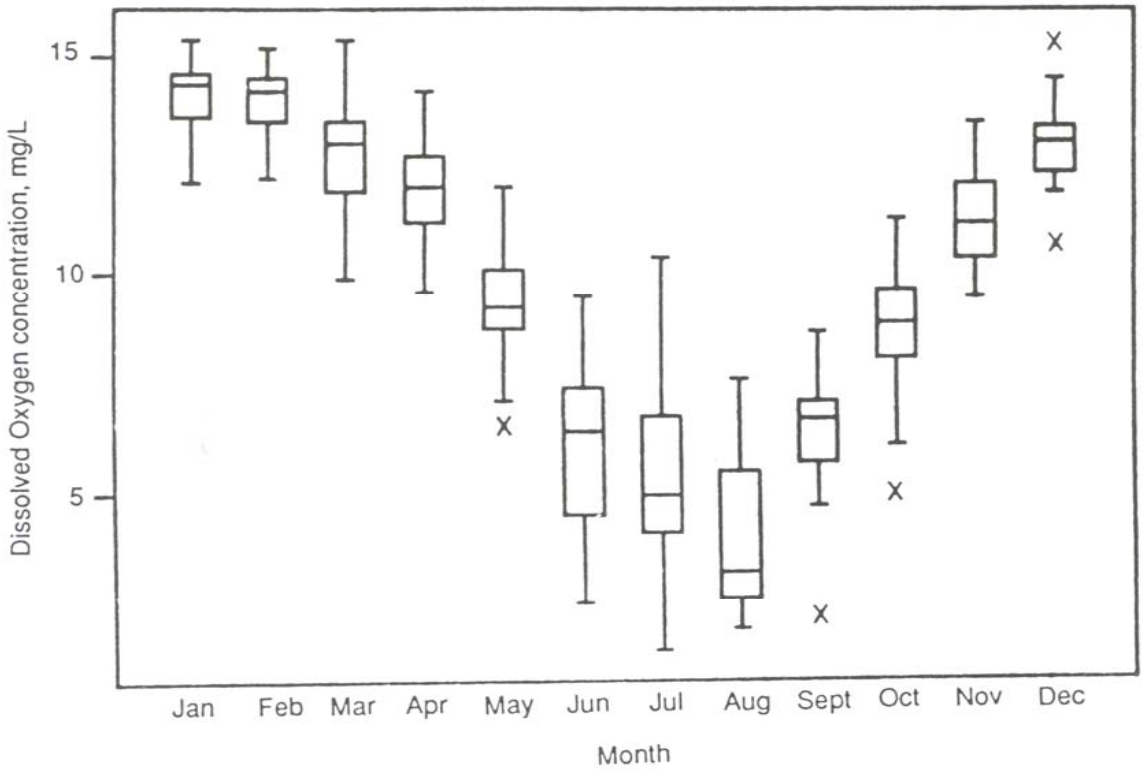
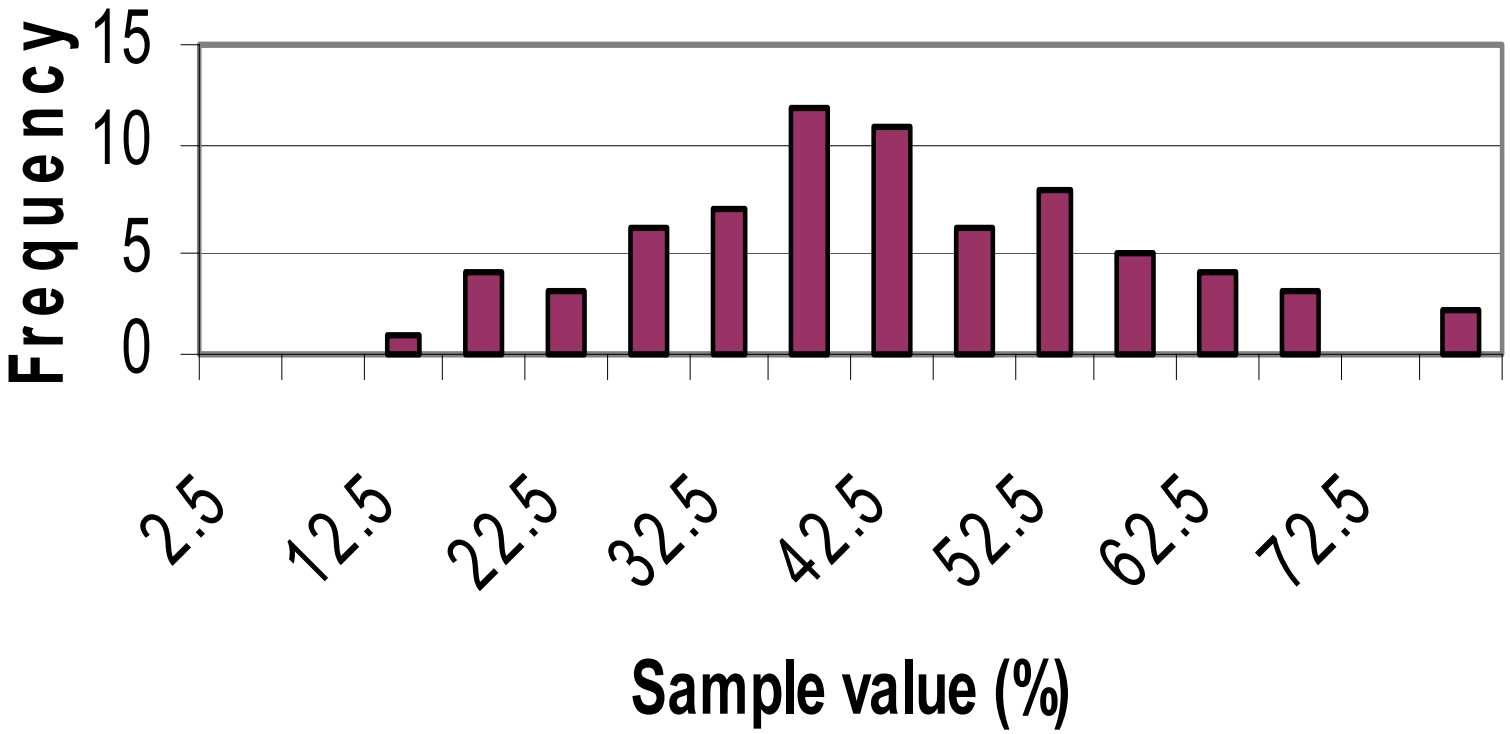


FIGURE 17.2.5 Side-by-side box plots of dissolved oxygen concentrations (in mg/L) measured at Conowingo Dam on the Susquehanna River, 1979–1989, by month of the year.

Maidment, D., 1993, Handbook of Hydrology

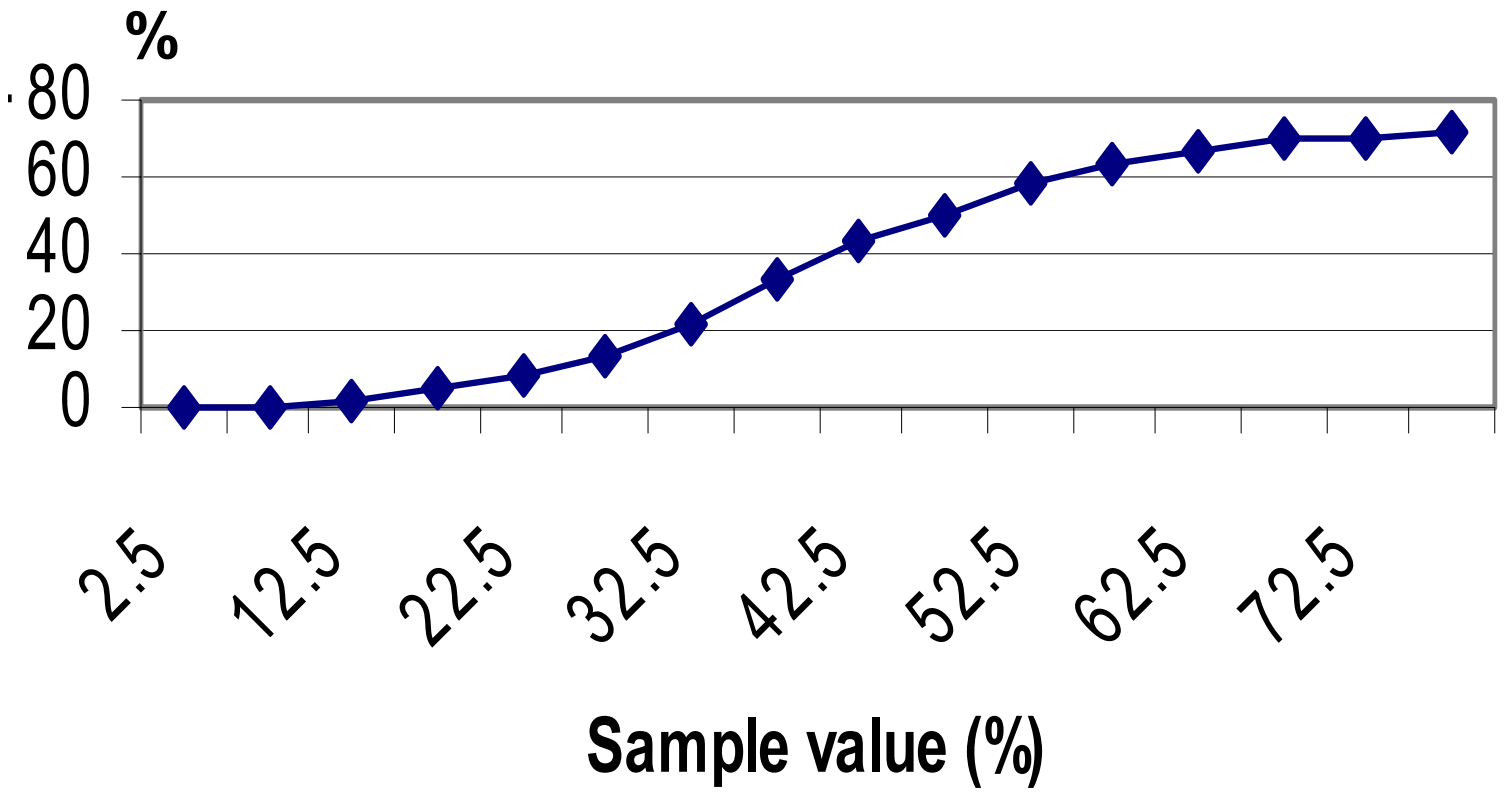
Graphical Display of Data

The Histogram



Graphical Display of Data

Cumulative Frequency Curve (Quantile Plot)



Graphical Display of Data

Probability Plot

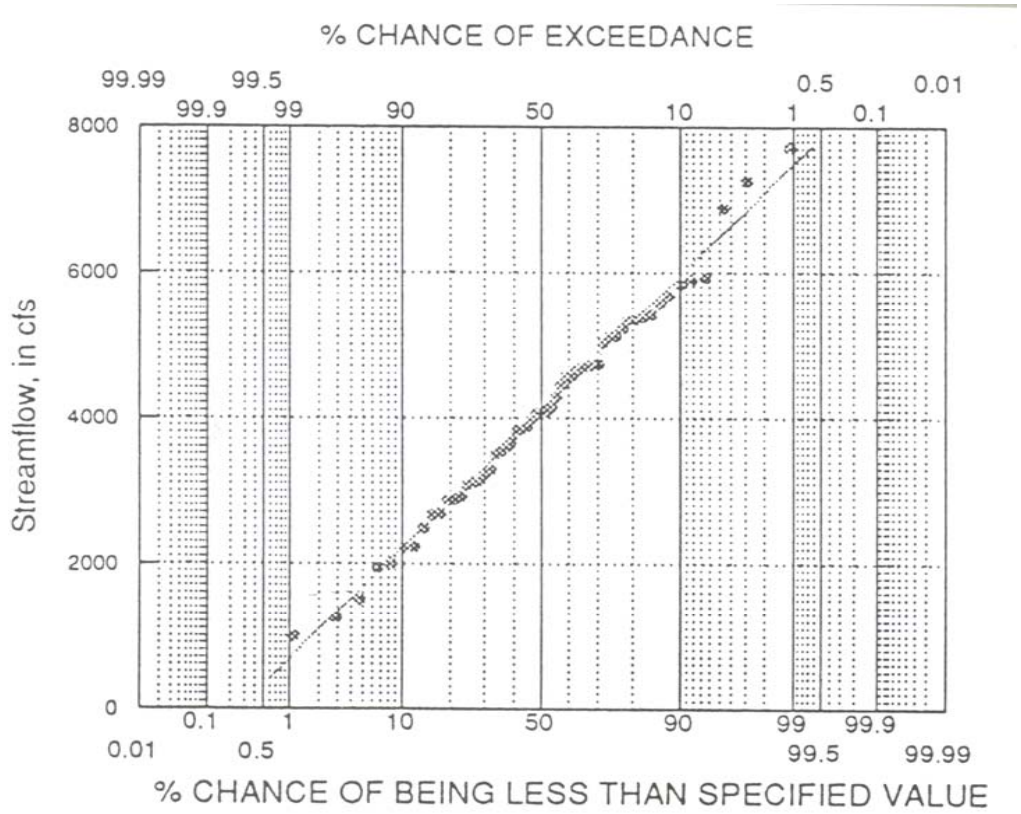


FIGURE 17.2.3 Probability plot of the annual stream flow, Licking River at Catawba, Ky., 1929–1983, on normal probability paper.

Maidment, D., 1993, Handbook of Hydrology

What is Probability?

- **Probability** is a measure of how likely (probable) specific observations of a variable may or may not occur.

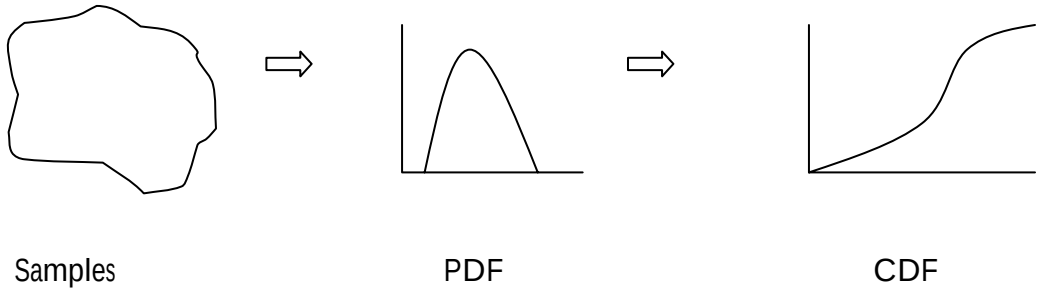
Example:

» P(porosity <= 0.15) = 10% or 0.10

» P(porosity > 0.23) = 82% or 0.82

How to Represent Probability?

1- Frequency distributions (e.g. cumulative frequency dist., quantile plot, probability plot … etc.)



How to Represent Probability?

2- Expectations

If P (porosity > 0.15) = 0.2 or (20% of the total volume of aquifer) and the aquifer volume is 50 million cubic meter,

- What is the expected volume of part of the aquifer that has porosity > 0.15?
- Expected volume of aquifer with porosity > 0.15 = 50 million m^3 X 0.2 = 10,000,000 m^3
- What is the expected amount of water stored in that volume?
- Expected amount of stored water = 10,000,000 m^3 X 0.15 = 1,500,000 m^3

Q: Is it feasible to utilize this aquifer or not?

Bivariate Statistics

- Bivariate statistics deals with the organization, presentation, and summary of data of **TWO** variables.
  - In geochemical analysis ➔ relationship between presence of a chemical element and water quality?
  - In carbonate rocks ➔ relationship between porosity and magnesium content?

- Bivariate statistics are represented by:
  - Summary statistics
  - Graphical plots

Summary Statistics

Correlation Coefficient

Quantitative measures of the linear **relationship** between two variables.

Covariance

$$COV(u,v) = S_{uv} = \frac{1}{n-1}\sum_{i=1}^{n}(u_i - m_u)(v_i - m_v)$$

Correlation Coefficient

$$r = \rho = \frac{S_{uv}}{S_u S_v}$$

Summary Statistics

Correlation Coefficient

- Its value range is between **–1** (completely negative ➔ one variable increases at the expense of the other!) and **+1** (completely positive ➔ both variables increase and decrease in the same way!)

- It is **sensitive** to outliers

- The absence of linear relationship does not mean that there is no relationship between the two variables ➔ **Non-linear relationship** might exist!

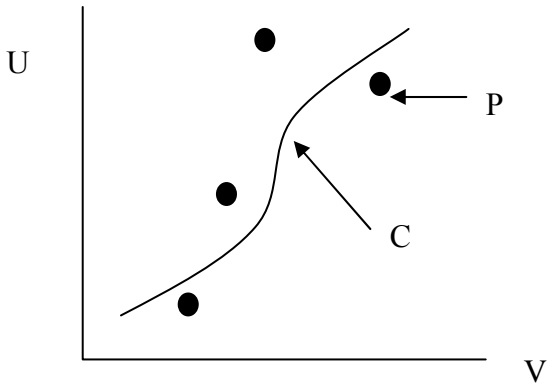
Regression Analysis

Def.: The process of **fitting a curve** through the cloud of point of a **scatterplot.**

The Process ➔ The Method of Least Squares

Of all curves approximating a given set of data points, the curve having the property that $D_1^2 + D_2^2 + \ldots\ldots\ldots + D_n^2$ is a minimum is called a **Best-Fitting Curve** or **Least-Squares Curve**.

D = The deviation or residual error between curve C and each point P.



Why regression analysis?
It helps in predicting one variable from the other (e.g. **censored** or **undetected** data)

Linear Regression

A mathematical expression that models the linear relationship between two variables (for example u and v).

$$u = a + bv$$

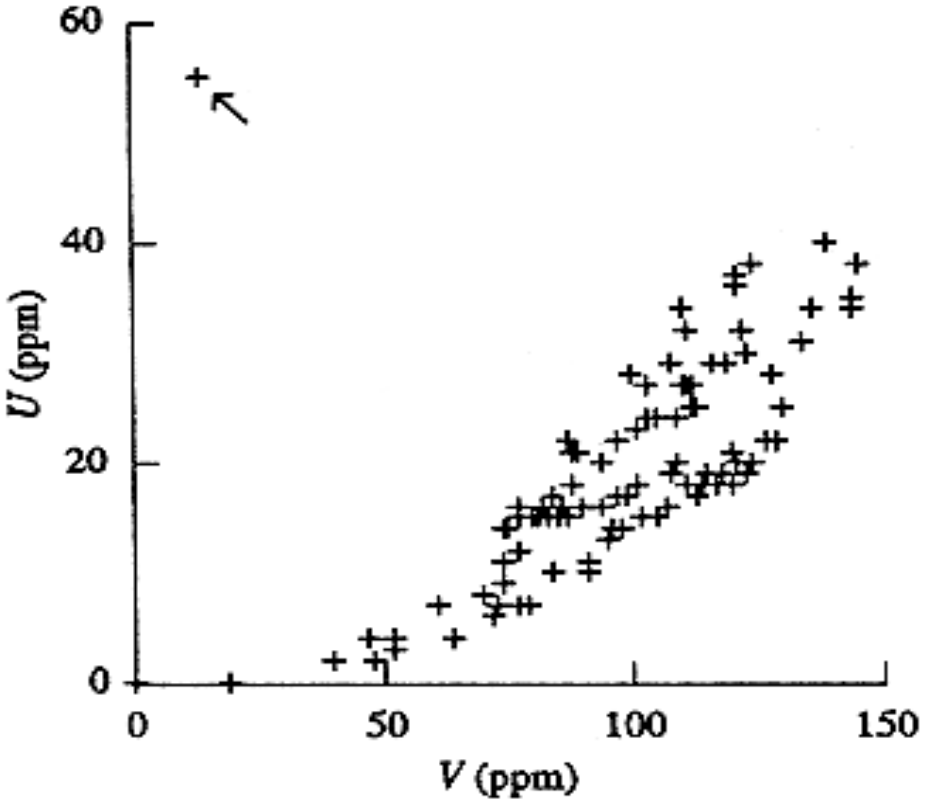
u = The **dependent** or regressed or predicted variable
v = The **independent** or regressor or predictor variable
a = The **intercept** of the line on the u axis

$$a = m_u - bm_v$$

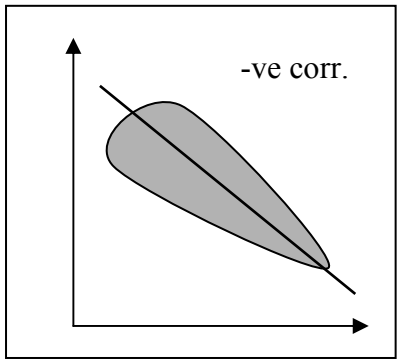
b = The **slope** of the line

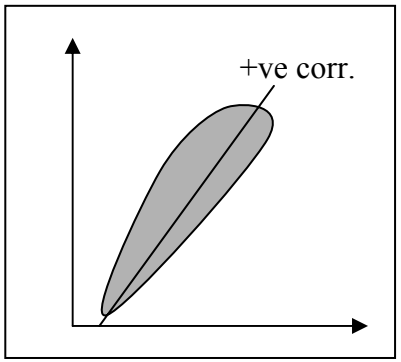
$$b = \rho \frac{S_u}{S_v}$$

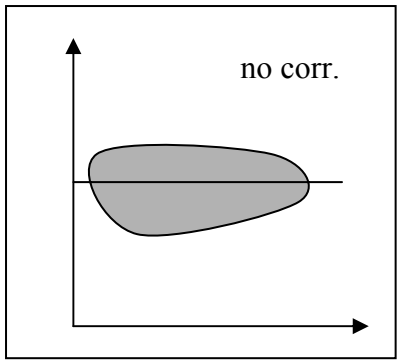
Graphical Display
Scatterplot

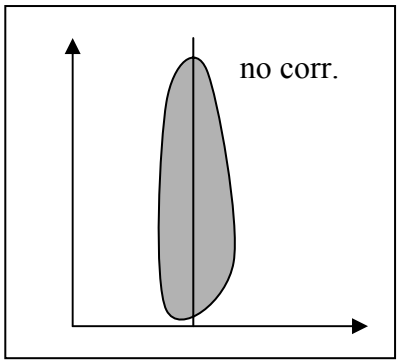


Positively or Negatively correlated, or Uncorrelated Variables









Decision-Making Practices

1) Estimation of the required number of samples

In order to **increase precision**, the size of confidence interval **(C.I.)** around a statistical parameter (say the mean) should be **decreased**.

To apply this, a **pilot sample** n_1 of a pre-investigation project must be used to conduct a preliminary statistical analysis.

With the help of the Student's **t-distribution table**, the **required number of samples** can be estimated **by minimizing the half of the width (d) of C.I.**

To minimize **d**, **n** must be increased.

If the pilot number of samples is n_1, the **approximate total required number of samples, n,** is derived as follows:

$$d = t_{(\frac{\alpha}{2},\nu_1)} \frac{s_1}{\sqrt{n}}$$

$$n = \frac{\left\{ t^2_{(\frac{\alpha}{2},\nu_1)} s_1^2 \right\}}{d^2}$$

Decision-Making Practices

2) Comparison of two means

It is revealed by testing hypothesis using the **t-distribution**

It is a useful test to understand if a **significant difference** occurs between **the means of two populations:**

- Are hydraulic conductivities similar in the upper & lower parts of an aquifer?
- Has the concentration of a pollutant declined in part of an aquifer as a result of installing a treatment plant?
- Does the concentration of a contaminant increase in an aquifer due to seasonal agricultural activities?

The test is performed by:

1) Computing **estimated t-statistics (t_{est})**

$$t_{est} = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S = \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}$$

2) Predicting the **critical t-value** (t_c)from the **Student's t-table**

$$t_c = t_{(\frac{\alpha}{2}, \nu)}$$

3) Testing the hypothesis:

a) If $t_{est} < t_c$ ➔ Accept **Null (No Difference) Hypothesis H_o** (i.e. the two means are equal)

b) If $t_{est} > t_c$ ➔ Reject **Null Hypothesis H_o** and accept the **Alternative Hypothesis H_a** (i.e. the two means are not equal)

Decision-Making Practices

3) Comparison of two variances

It is revealed by testing hypothesis using the **F-distribution** or Fisher's distribution (After R. A. Fisher who discovered it).

It is a useful test to understand if a **significant difference** occurs between **the variances of two populations**

The test is performed by:

1) Computing estimated F-statistics (F_{est})

$$F_{est} = \frac{S_1^2}{S_2^2} \qquad S_1 > S_2$$

2) Predicting the critical F-value (F_c)from the F-table

$$F_c = F(\alpha, \upsilon_1, \upsilon_2)$$

3) Testing the hypothesis:

a- If $F_{est} < F_c$ ➔ Accept Null (No Difference) Hypothesis H_o (i.e. the two variances are equal)

b- If $F_{est} > F_c$ ➔ Reject Null Hypothesis H_o and accept the Alternative Hypothesis H_a (i.e. the two variances are not equal)

Decision-Making Practices
Example for t- & F- tests

	Data # 1		Data # 2	
		NO3		NO3
	Aug	ppm	May	ppm
	Arj001	70	WS01	119.8
	Arj019	47	WS02	186.9
	Arj020	45	WS3	75.2
	Arj002	46	WS10	89.3
	Arj003	50	WS4	600.45
	Arj022	46	WS06	106.4
	Arj035	52	WS07	107.4
	Arj023	80	WS08	142.2
	Arj004	46	WS09	83.2
	Arj024	50	WS15	128.5
	Arj006	71	WS12	145.4
	Arj005	91	WS13	141
	Arj007	54	WS11	340.8
	Arj026	54	WS14	101
	Arj027	66		
	Arj009	53		
	Arj011	66		
	Arj014	50		
	Arj028	73		
	Arj010	77		
	Arj029	63		
	Arj013	68		
	Arj036	19		
	Arj016	141		
	Arj017	46		
	Arj032	64		
	Arj018	33		
	Arj033	68		
	Arj034	48		
count		29		14
mean		59.8966		169.111
variance		465.81		19765.7

F-test

F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	59.8966	169.1107143
Variance	465.81	19765.70468
Observations	29	14
df	28	13
F	**0.02357**	
P(F<=f) one-tail	6.3E-15	
F Critical one-tail	**0.47871**	

t-test

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	169.111	59.89655172
Variance	19765.7	465.8103448
Observations	14	29
Pooled Variance	6585.29	
Hypothesized Mean Difference	0	
df	41	
t Stat	**4.13543**	
P(T<=t) one-tail	8.6E-05	
t Critical one-tail	1.68288	
P(T<=t) two-tail	0.00017	
t Critical two-tail	**2.01954**	