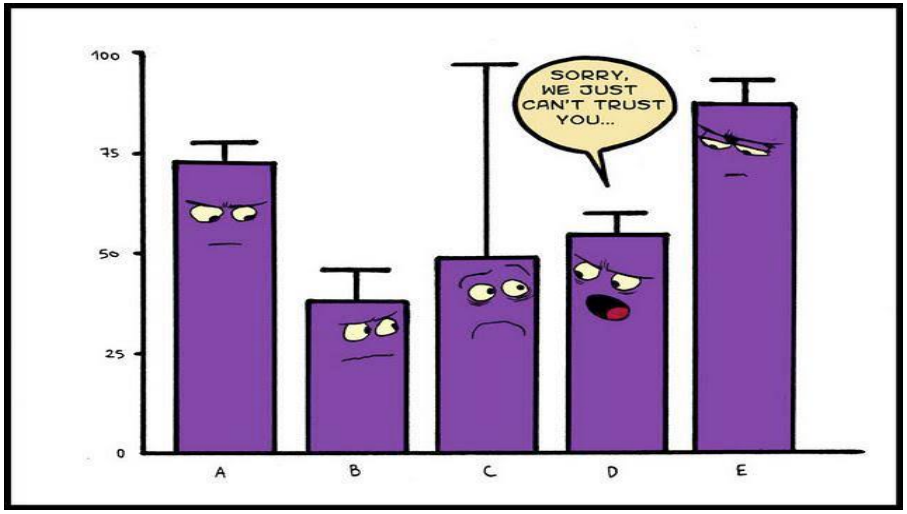
INTERNATIONAL GUEST LECTURE
Fakultas Teknik, Universitas Negeri Malang

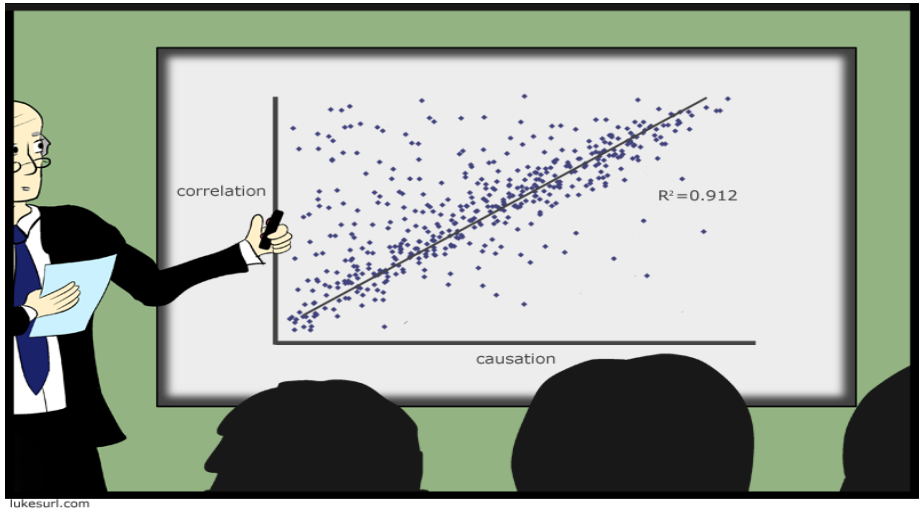
EXPERIMENTAL DATA ANALYSIS

Norazian Mohamed Noor, PhD

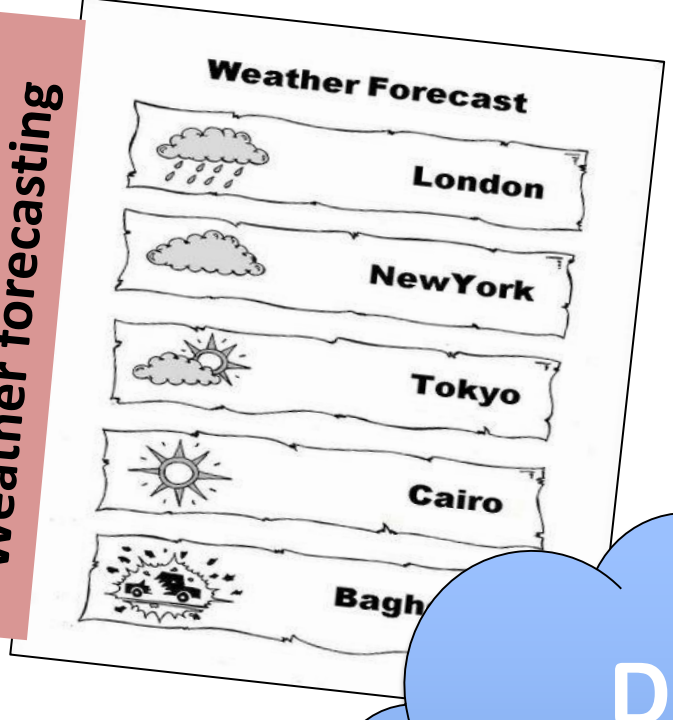
School of Environmental Engineering

Universiti Malaysia Perlis (UniMAP)



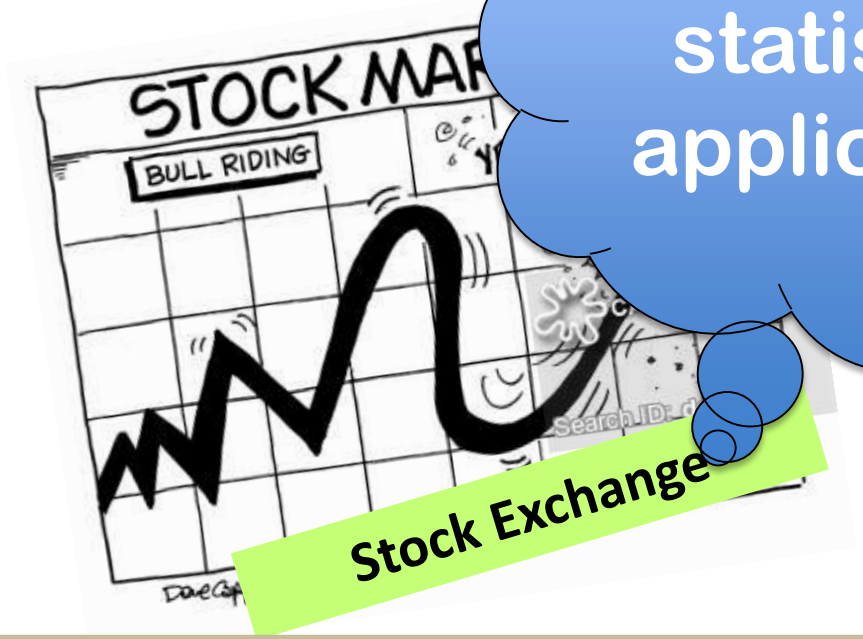


Weather forecasting



Classroom performances

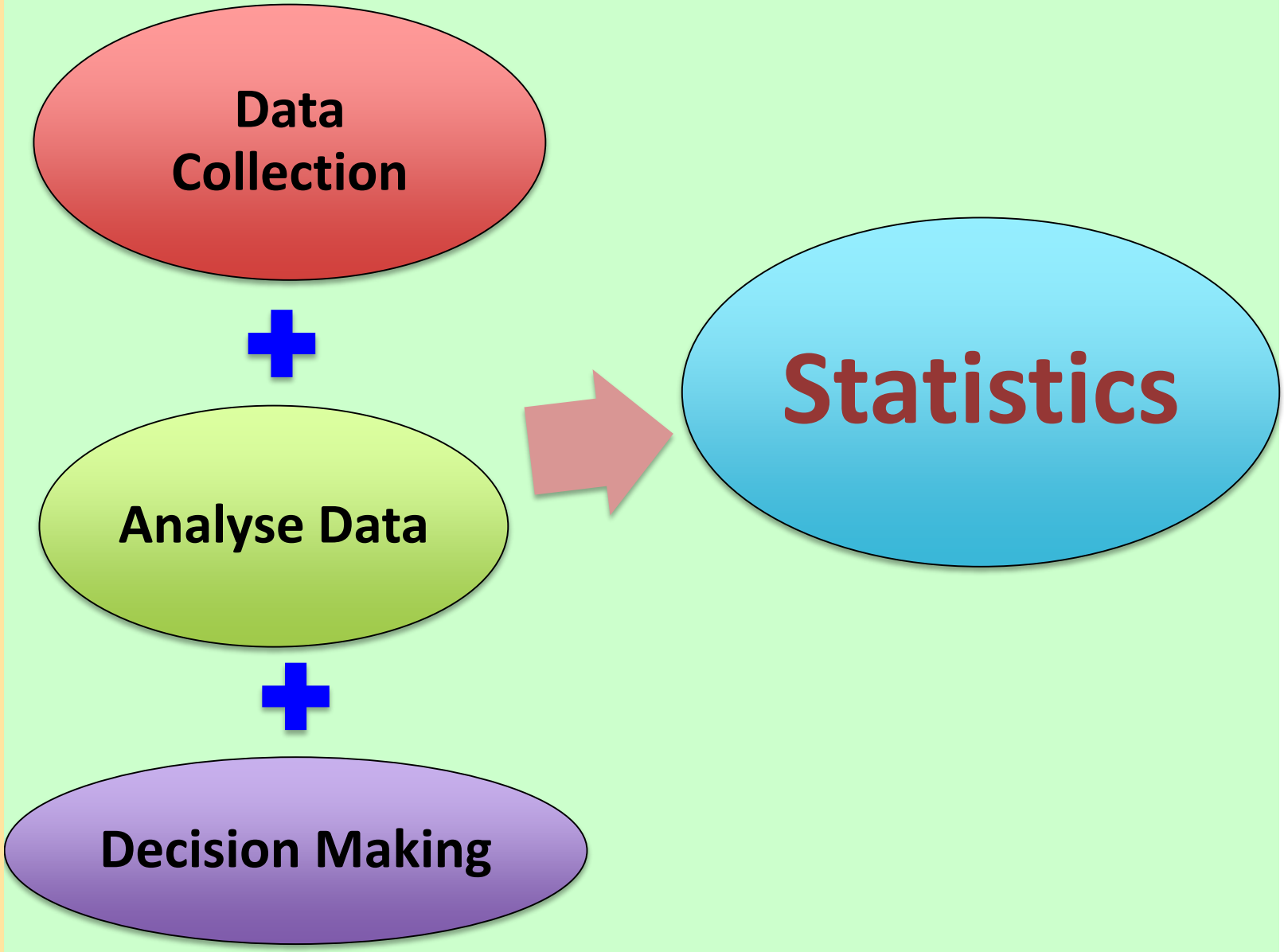
Daily statistics application



Stock Exchange

Income taxes

WHAT IS STATISTICS??

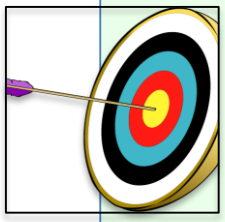


GENERAL QUANTITATIVE RESEARCH FLOWCHART

1. Review the project objectives and sampling design
2. Conduct a preliminary data review
3. Select the statistical method
4. Verify the assumptions of the statistical method
5. Draw conclusions from the data

1

- Review the project objectives and sampling design

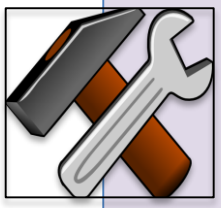


PURPOSE

* Review the objectives, the sampling design and the consistency of the data collection.

ACTIVITIES

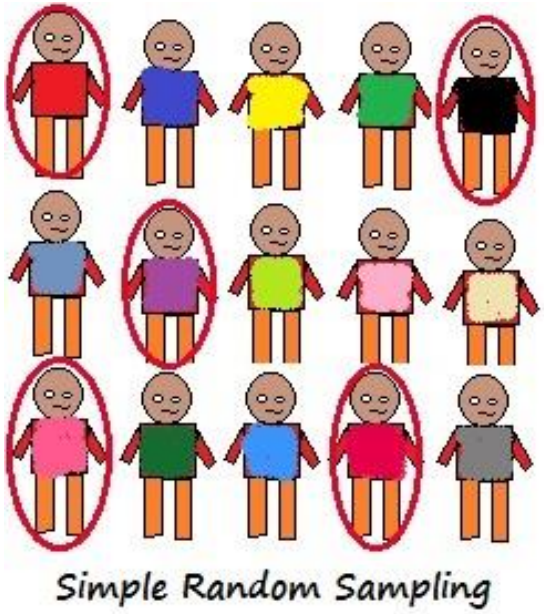
* Review Study Objectives
* Translate the objectives into statistical hypothesis
* Develops the limits on Decision Errors
* Review the sampling designs

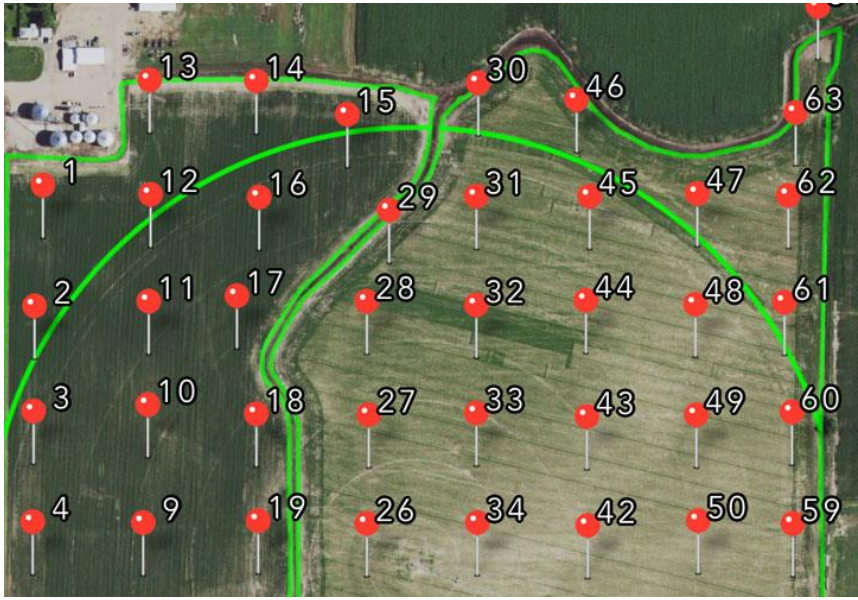


TOOLS

* Hypothesis statement
*Sampling concept

Examples of some samplings method





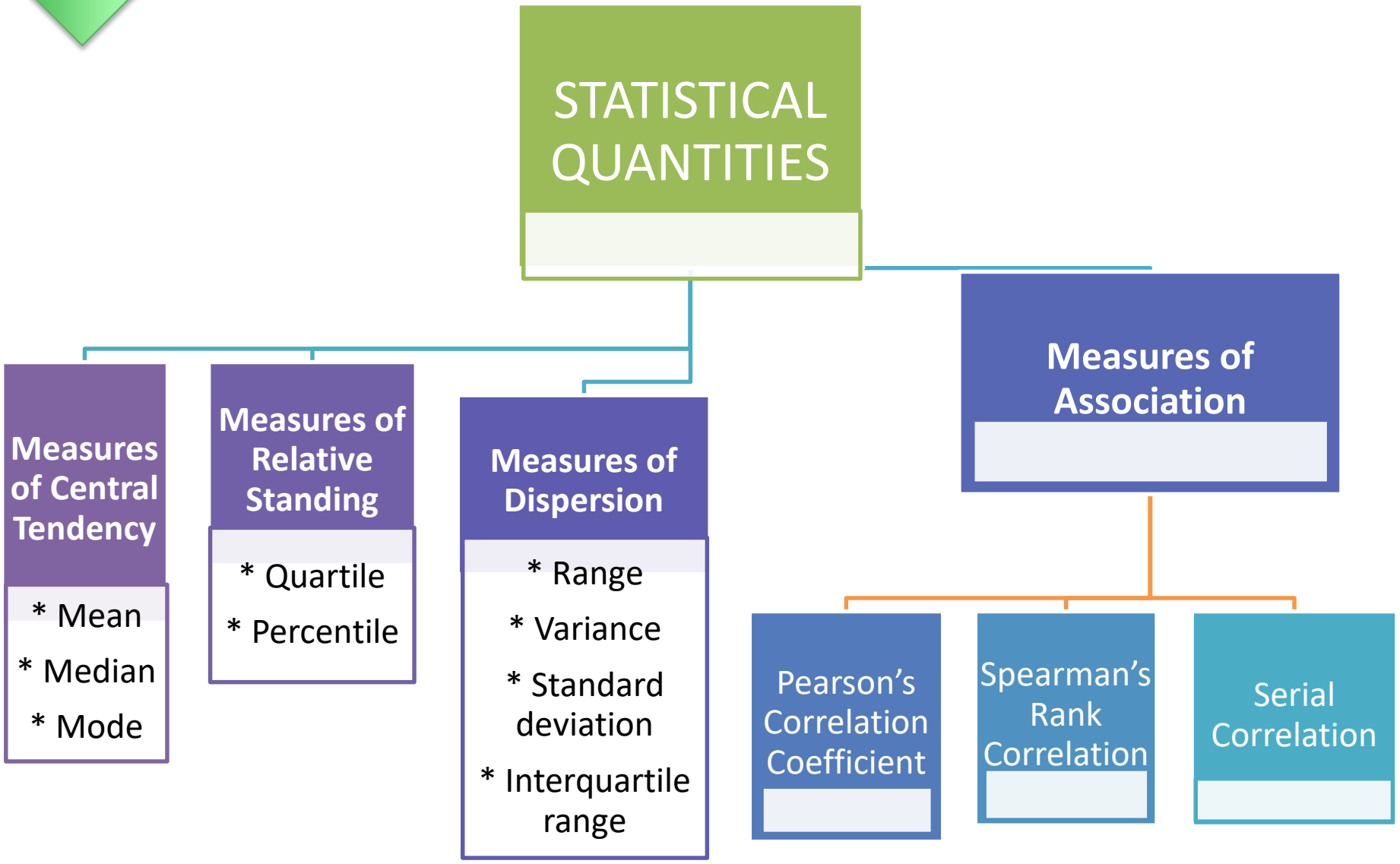
Grid sampling

GENERAL QUANTITATIVE RESEARCH FLOWCHART

1. • Review the project objectives and sampling design
2. • Conduct a preliminary data review
3. • Select the statistical method
4. • Verify the assumptions of the statistical method
5. • Draw conclusions from the data

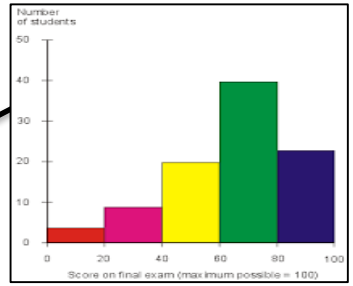
2

- Conduct a preliminary data review



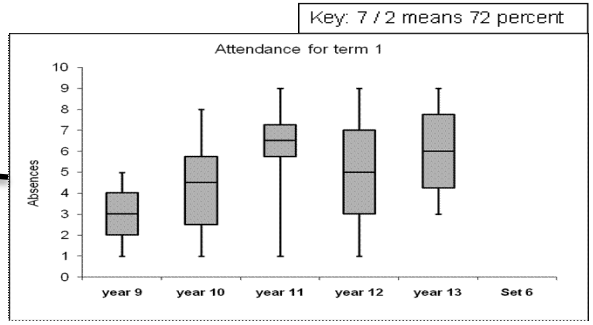
GRAPHICAL PRESENTATIONS

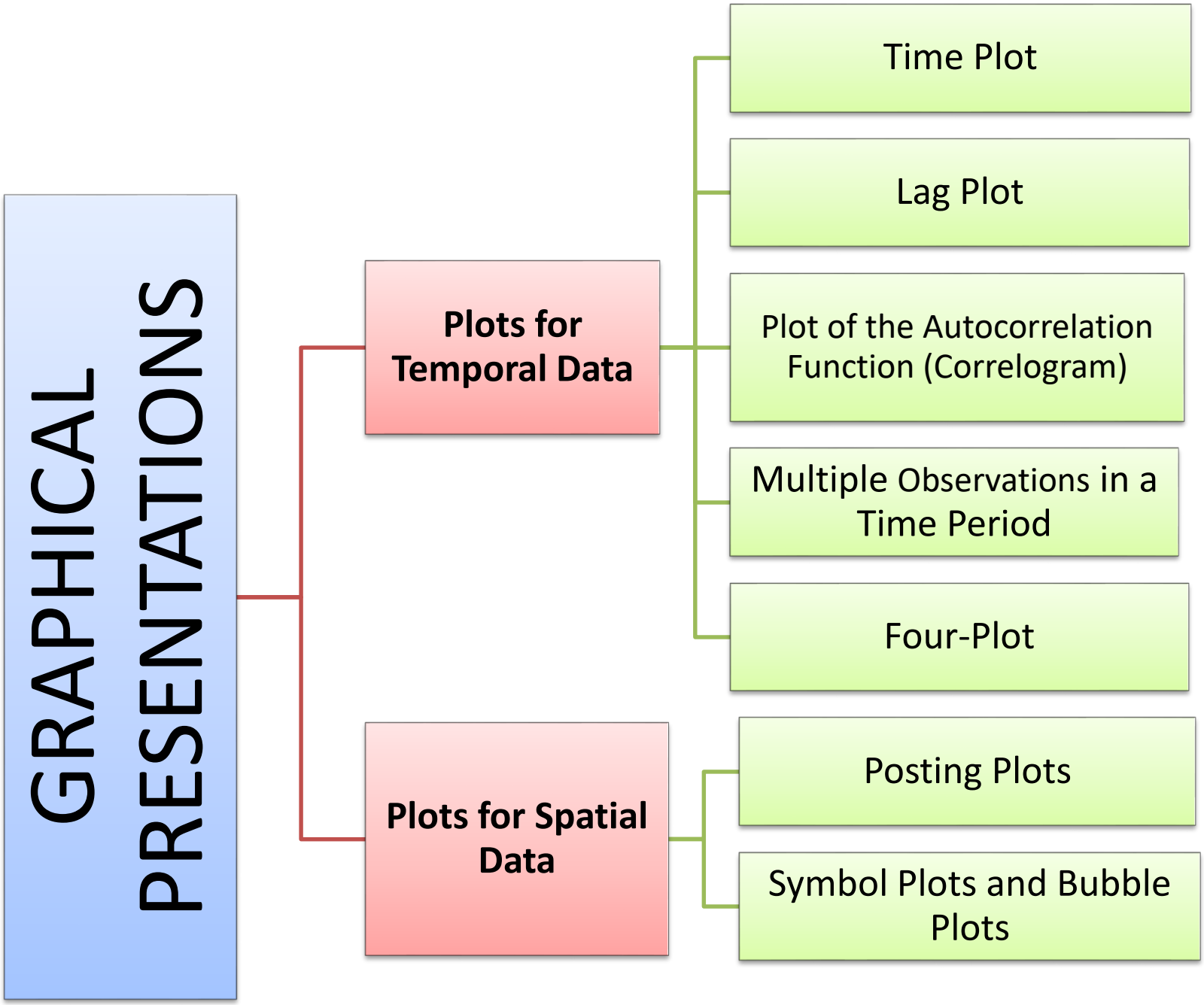
- Histogram
- Stem-to-leaf Plot
- Box-and-Whiskers Plot
- Quantile Plot and Ranked Data Plots
- Plots for Two or More Variables
  - Scatterplot
  - Extensions of the Scatterplot
  - Empirical Quantile-Quantile Plot



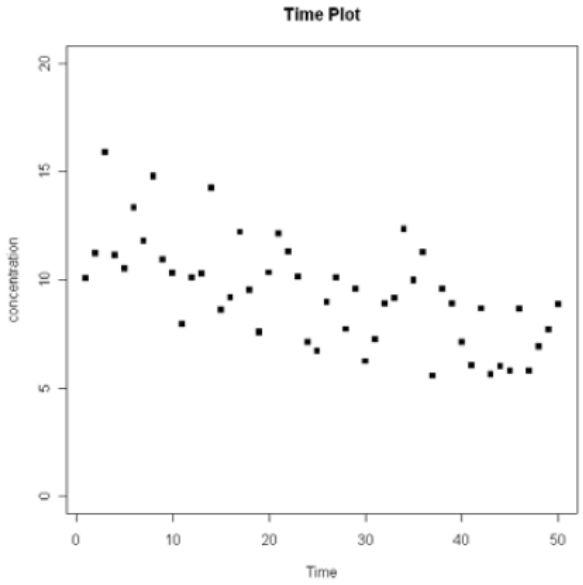
Grades on a Science Test

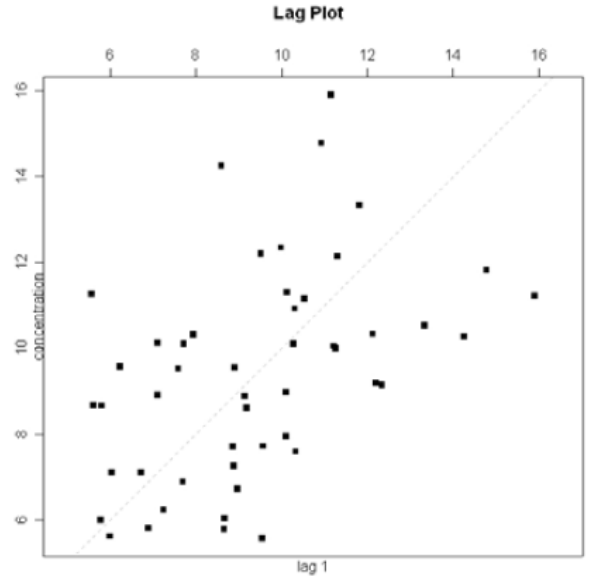
Stem	Leaf
7	2 2 4 5 6 9
8	1 4 5 7 7 9
9	0 1 3 5 8
10	0 0

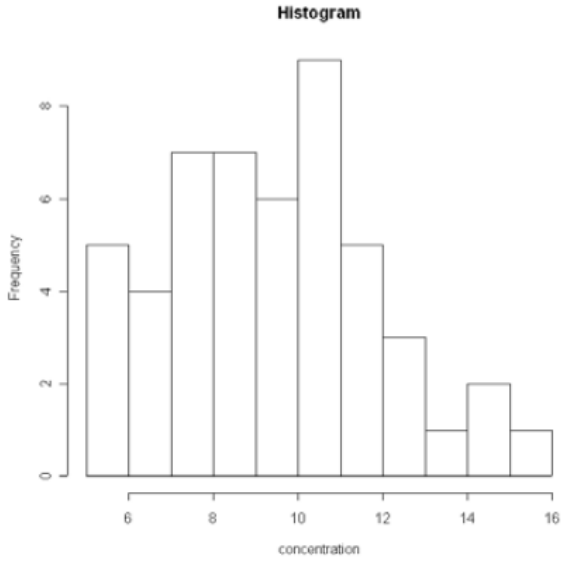


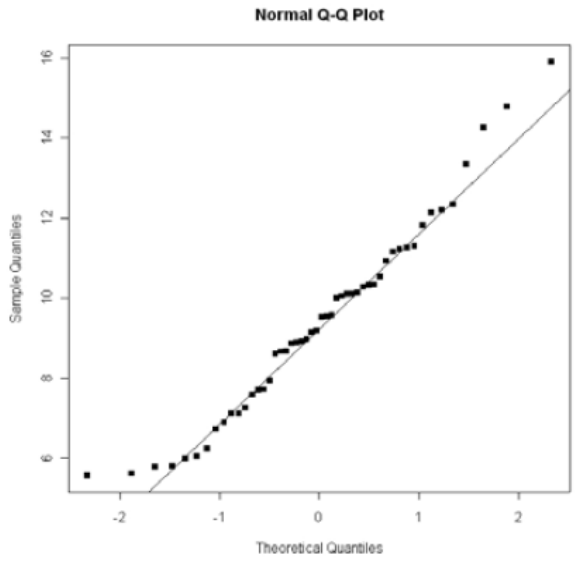

GRAPHICAL PRESENTATIONS
Plots for Temporal Data
Time Plot
Lag Plot
Plot of the Autocorrelation Function (Correlogram)
Multiple Observations in a Time Period
Four-Plot
Plots for Spatial Data
Posting Plots
Symbol Plots and Bubble Plots

EXAMPLES OF GRAPHICAL PRESENTATION FOR A TEMPORAL DATA



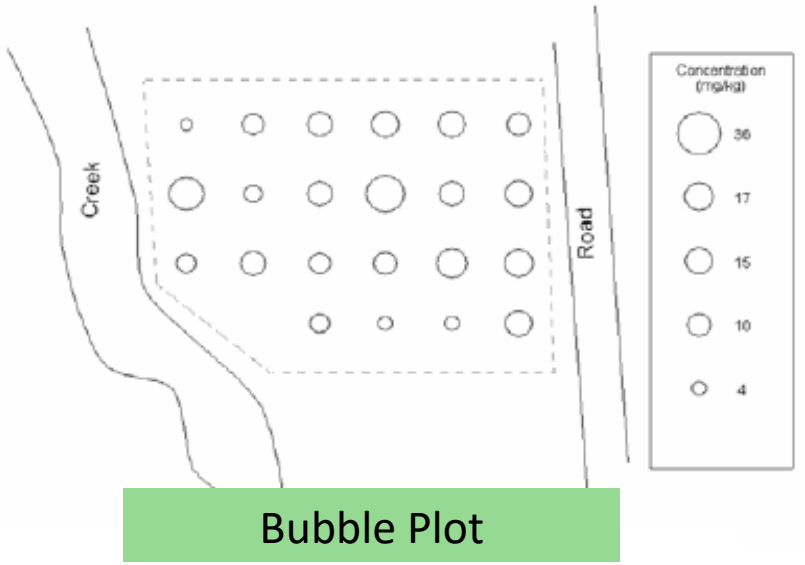




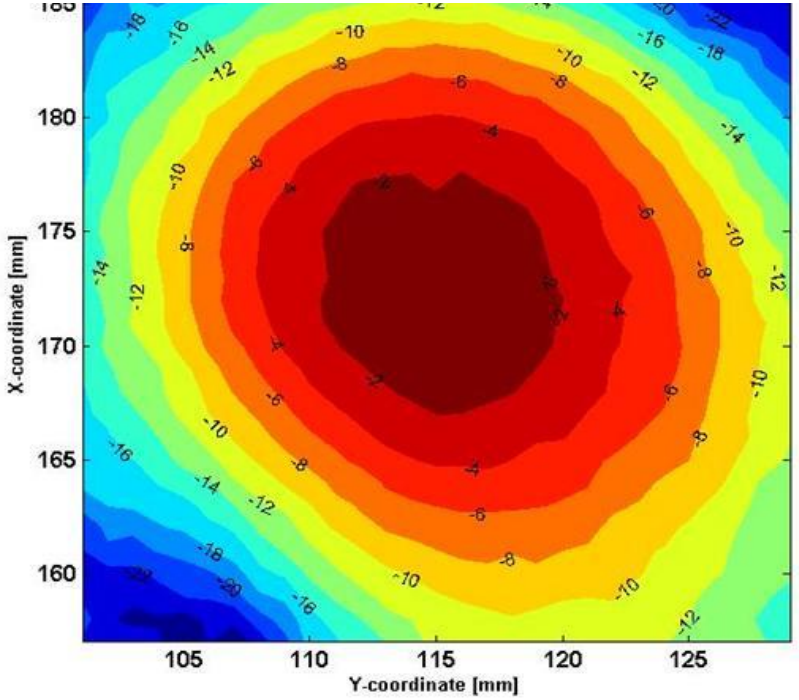


EXAMPLES OF GRAPHICAL PRESENTATION FOR A SPATIAL DATA

Symbol Plot



Bubble Plot



Contour plot

GENERAL QUANTITATIVE RESEARCH FLOWCHART

1. • Review the project objectives and sampling design
2. • Conduct a preliminary data review
3. • Select the statistical method
4. • Verify the assumptions of the statistical method
5. • Draw conclusions from the data

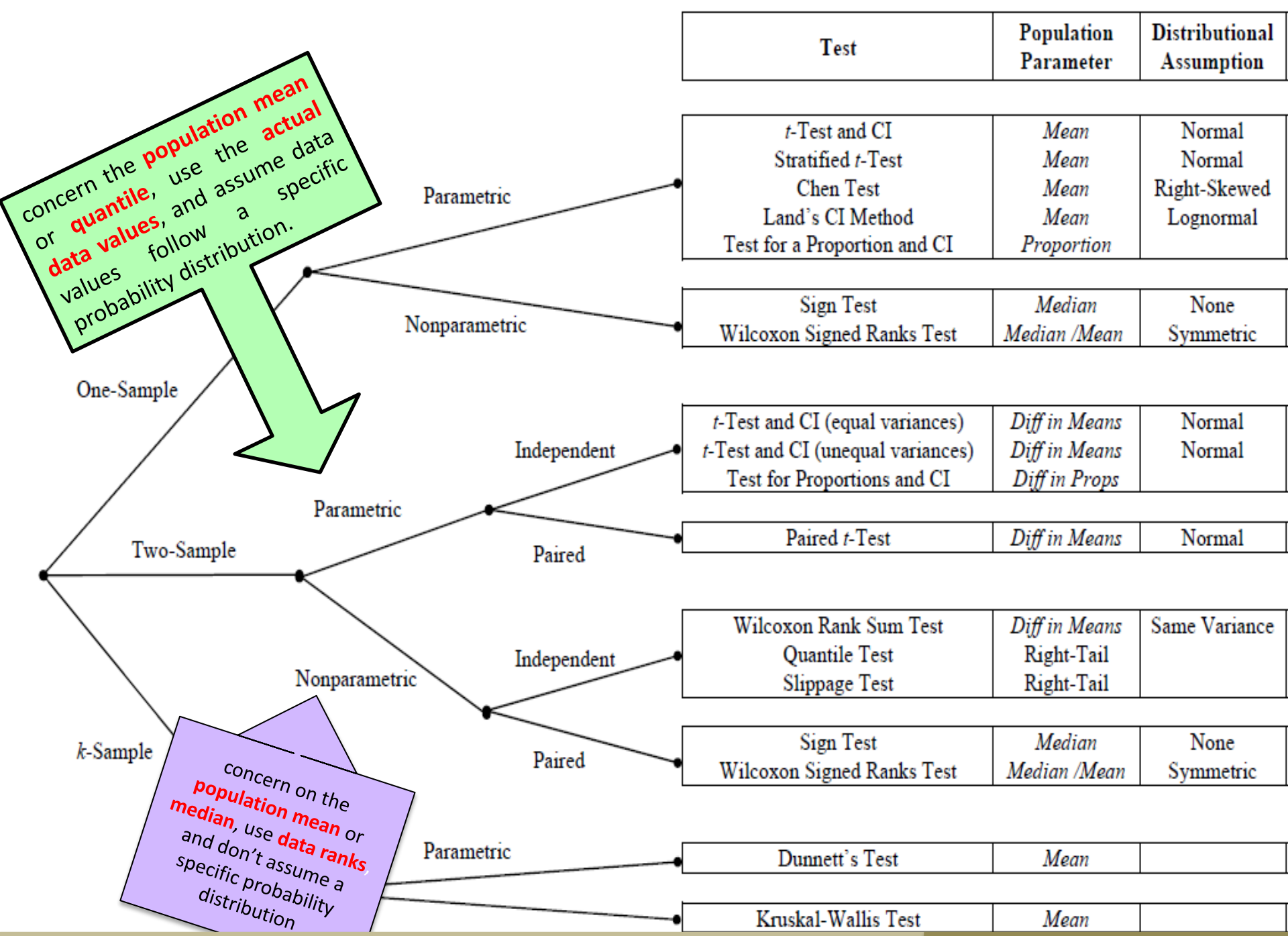
3 • Select the statistical method

- There are two important outputs from this step:

  1. **the chosen method, and**
  2. **the assumptions underlying the method.**

- If a **particular statistical procedure has been specified** either during the determination of the objective; or the particular program or study, the analyst should use the results of the preliminary data review to **determine if it is appropriate for the data collected.**

- If a particular procedure has **not been specified**, then the analyst should **select one based upon the data user's objectives and the preliminary data review**.



Test	Population Parameter	Distributional Assumption
t-Test and CI	*Mean*	Normal
Stratified *t*-Test	*Mean*	Normal
Chen Test	*Mean*	Right-Skewed
Land's CI Method	*Mean*	Lognormal
Test for a Proportion and CI	*Proportion*	
Sign Test	*Median*	None
Wilcoxon Signed Ranks Test	*Median /Mean*	Symmetric
t-Test and CI (equal variances)	*Diff in Means*	Normal
t-Test and CI (unequal variances)	*Diff in Means*	Normal
Test for Proportions and CI	*Diff in Props*	
Paired *t*-Test	*Diff in Means*	Normal
Wilcoxon Rank Sum Test	*Diff in Means*	Same Variance
Quantile Test	Right-Tail	
Slippage Test	Right-Tail	
Sign Test	*Median*	None
Wilcoxon Signed Ranks Test	*Median /Mean*	Symmetric
Dunnett's Test	*Mean*	
Kruskal-Wallis Test	*Mean*	

GENERAL QUANTITATIVE RESEARCH FLOWCHART

1. Review the project objectives and sampling design
2. Conduct a preliminary data review
3. Select the statistical method
4. Verify the assumptions of the statistical method
5. Draw conclusions from the data

4 • Verify the assumptions of the statistical method

- Assess the **validity** of the **statistical method** chosen by examining its **underlying assumptions** or determine that the data support the underlying assumptions necessary for the selected method, or if a different statistical method should be used.

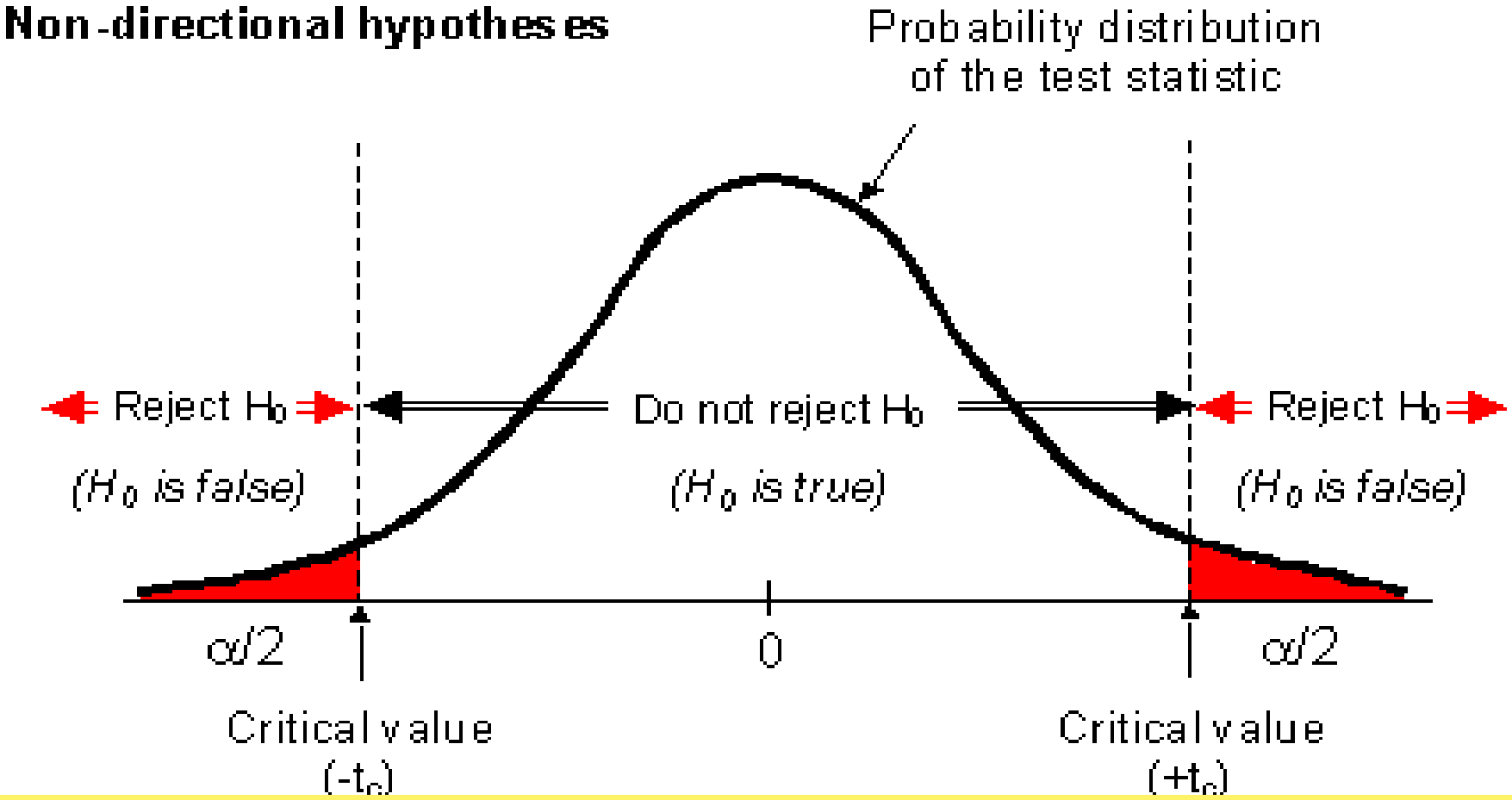
t-Test	Paired t-Test (Equal variances)	ANOVA
Sample is random	The two populations are independent	Different populations are random and independent.
The sample size is large	Both are approximately normally distributed	The populations is normally distributed
Population mean is known	The variances of both populations are approximately equal	The variances of both populations are approximately equal

GENERAL QUANTITATIVE RESEARCH FLOWCHART

1. Review the project objectives and sampling design
2. Conduct a preliminary data review
3. Select the statistical method
4. Verify the assumptions of the statistical method
5. Draw conclusions from the data

5

- Draw conclusions from the data



REJECTION RULE

(p-value Approach)

FORMULATION OF HYPOTHESIS: 2 INDEPENDENT/PAIRED MEAN

Mean Comparison	H_o & H_A	Types of Test
$\mu_2 > \mu_1$	$H_A: \mu_2 > \mu_1$ $H_o: \mu_2 \leq \mu_1$	1 right-tailed
$\mu_2 < \mu_1$	$H_A: \mu_2 < \mu_1$ $H_o: \mu_2 \geq \mu_1$	1 left-tailed
$\mu_2 = \mu_1$	$H_A: \mu_2 < \mu_1$ $H_o: \mu_2 \geq \mu_1$	2-tailed
$\mu_2 \geq \mu_1$	$H_A: \mu_2 < \mu_1$ $H_o: \mu_2 \geq \mu_1$	1 left-tailed
$\mu_2 \leq \mu_1$	$H_A: \mu_2 > \mu_1$ $H_o: \mu_2 \leq \mu_1$	1 right-tailed
$\mu_2 \neq \mu_1$	$H_A: \mu_2 \neq \mu_1$ $H_o: \mu_2 = \mu_1$	2-tailed

REJECTION RULE
(p-value Approach)

Comparison of p-value and α	Test Decision	Comment on Decision	How much Evidence
$p > 0.10$	Fail to reject H_o	Not significant	None
$0.05 < p \leq 0.10$	(Might) Reject H_o	Barely significant	Some/little
$0.01 < p \leq 0.05$	Reject H_o	Significant	Enough/ sufficient
$p \leq 0.01$	(Definitely) Reject H_o	Highly significant	Strong/ overwhelming

"Data don't make any sense,
we will have to resort to statistics."

T-test, paired t-test, Analysis of Variance (ANOVA)

STATISTICAL APPROACH ON EXPERIMENTAL DATA ANALYSIS

Examples: ANOVA

Water sample were taken at 4 different locations in a river to determine whether the quantity of dissolved oxygen, a measure of water pollution, varied from one location to another.

Location 1 and 2 were selected above an industrial plant, one near the shore and other in midstream; location 3 was adjacent to the industrial water discharge for the plant and location 4 was slightly downriver in midstream.

Five water specimens were randomly selected at each location which the amount of dissolved oxygen were recorded.

Examples: ANOVA

Data

Location

1	2	3	4
5.9	6.3	4.8	6.0
6.1	6.6	4.3	6.2
6.3	6.4	5.0	6.1
6.1	6.4	4.7	5.8
6.0	6.5	5.1	5.7

Solutions: ANOVA

1. Hypothesis

$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$ (the mean dissolved O_2 contents are the same at all 4 locations)

H_1 : at least one μ_i differs from one μ_i (at least 1 location differs from the other)

2. Checking for assumptions

- Normality – Normal Q-Q plot/ Shapiro-Wilks Test
- Equal variance – Levene's Test

3. Statistical Method – ANOVA

4. Decision making - Whether to Reject H_o or do not reject H_o.

Solutions: ANOVA

DESCRIPTIVE STATISTICS

Descriptives

DISSOLVED_O2

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
Location 1		5	6.080	.1483	.0663	5.896	6.264	5.9	6.3	
Location 2		5	6.440	.1140	.0510	6.298	6.582	6.3	6.6	
Location 3		5	4.780	.3114	.1393	4.393	5.167	4.3	5.1	
Location 4		5	5.960	.2074	.0927	5.703	6.217	5.7	6.2	
Total		20	5.815	.6675	.1493	5.503	6.127	4.3	6.6	
Model	Fixed Effects			.2092	.0468	5.716	5.914			
	Random Effects				.3598	4.670	6.960			.5090

NORMALITY ASSUMPTION

Normality Test

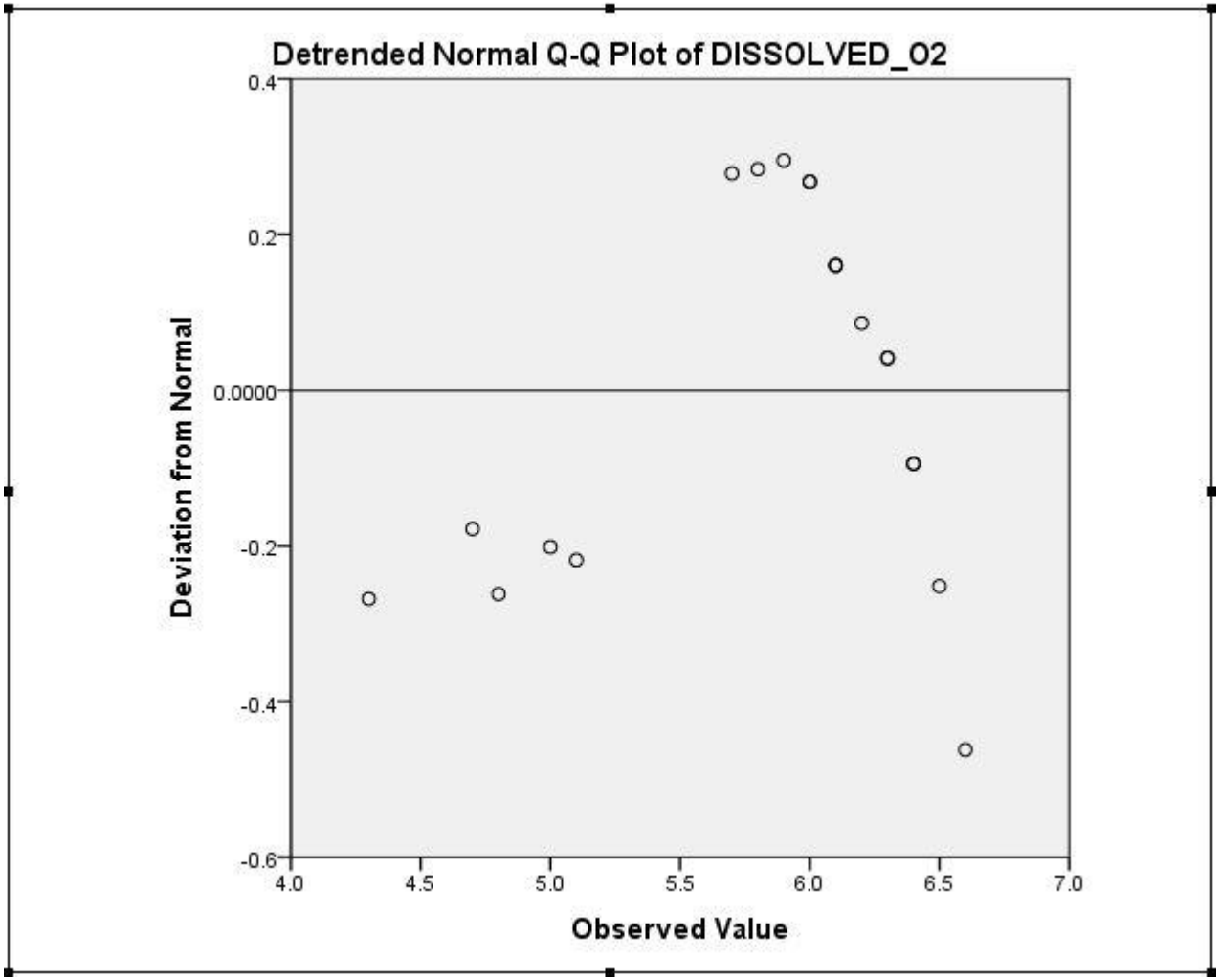
	Kolmogorov-Smimov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	Df	Sig.
DISSOLVED_O2	.091	20	.200*	.969	20	.725

*This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Since the p-value > 0.05, then the distribution is normal

Solutions: ANOVA



Solutions: ANOVA

EQUAL VARIANCE ASSUMPTION

Test of Homogeneity of Variances

DISSOLVED_O2

Levene Statistic	df1	df2	Sig.
1.449	3	16	.266

Since the p-value is 0.266 > 0.05, then the population have equal variances

ANOVA TABLE

ANOVA

DISSOLVED_O2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7.766	3	2.589	59.166	.000
Within Groups	.700	16	.044		
Total	8.466	19			

Since the p-value is 0.000 < 0.05, then there are differences in the mean

Solutions: ANOVA

- **Hypothesis:**

$$\mathbf{H_o}: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

(the mean of dissolved O_2 are the same for the 4 locations)

VS

$\mathbf{H_1}$: at least one μ_i differs from one μ_i.

(at least 1 location is different from the other)

- **Conclusion:**

Since $\alpha' = 0.000 < 0.05$, H_o is rejected at $\alpha = 0.05$. There are differences in mean dissolved O_2 content among the 4 locations.

Solutions: ANOVA

Post Hoc Tests

Multiple Comparisons

Dependent Variable: DISSOLVED_O2
Tukey HSD

(I) LOCATION	(J) LOCATION	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Location 1	Location 2	-.3600	.1323	.065	-.738	.018
	Location 3	1.3000*	.1323	.000	.922	1.678
	Location 4	.1200	.1323	.801	-.258	.498
Location 2	Location 1	.3600	.1323	.065	-.018	.738
	Location 3	1.6600*	.1323	.000	1.282	2.038
	Location 4	.4800*	.1323	.011	.102	.858
Location 3	Location 1	-1.3000*	.1323	.000	-1.678	-.922
	Location 2	-1.6600*	.1323	.000	-2.038	-1.282
	Location 4	-1.1800*	.1323	.000	-1.558	-.802
Location 4	Location 1	-.1200	.1323	.801	-.498	.258
	Location 2	-.4800*	.1323	.011	-.858	-.102
	Location 3	1.1800*	.1323	.000	.802	1.558

*. The mean difference is significant at the 0.05 level.

Solutions: ANOVA

CONCLUSION:

1. There is a **SIGNIFICANT DIFFERENCE** in the mean of dissolved O_2 in Location 3 (adjacent to the industrial water) in comparison with other locations.

1. There is **NO DIFFERENCE** in the mean of dissolved O_2 between:

- Location 1 (near the shore) and Location 2 (in the midstream).
- Location 1 (near the shore) and Location 4 (downriver in the midstream)