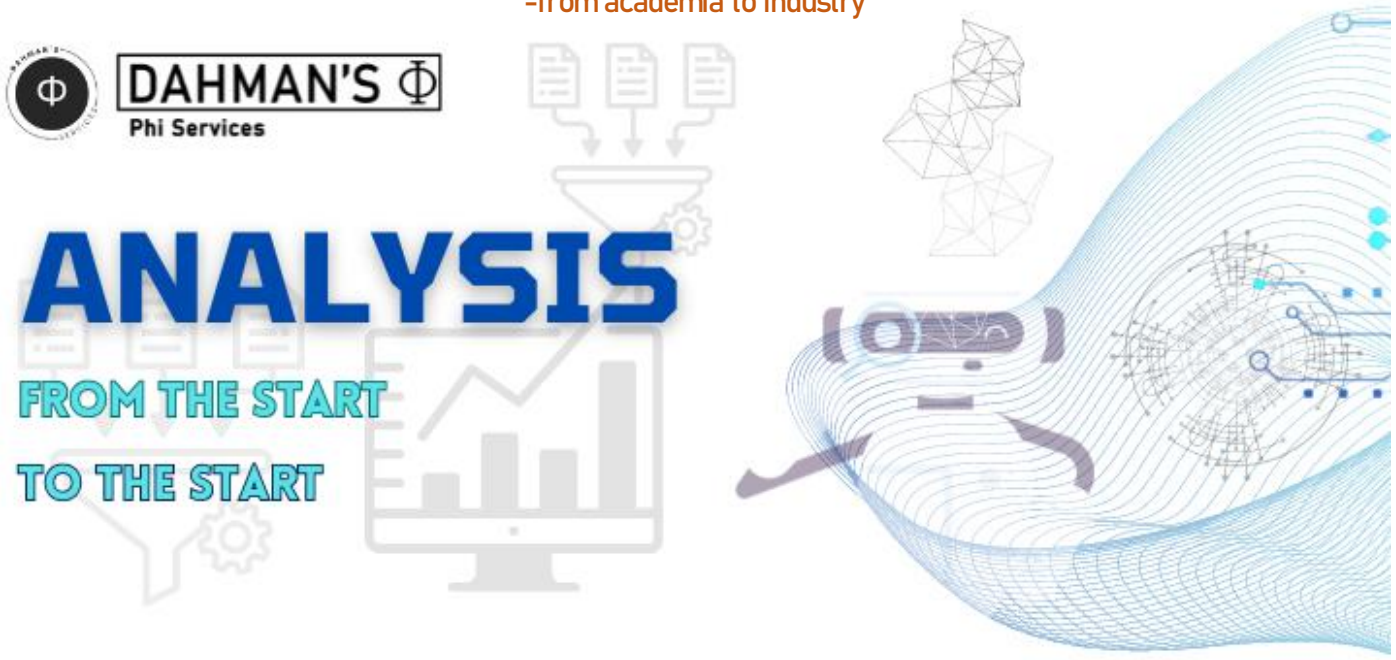
THE BIG BANG OF DATA SCIENCE

-from academia to industry

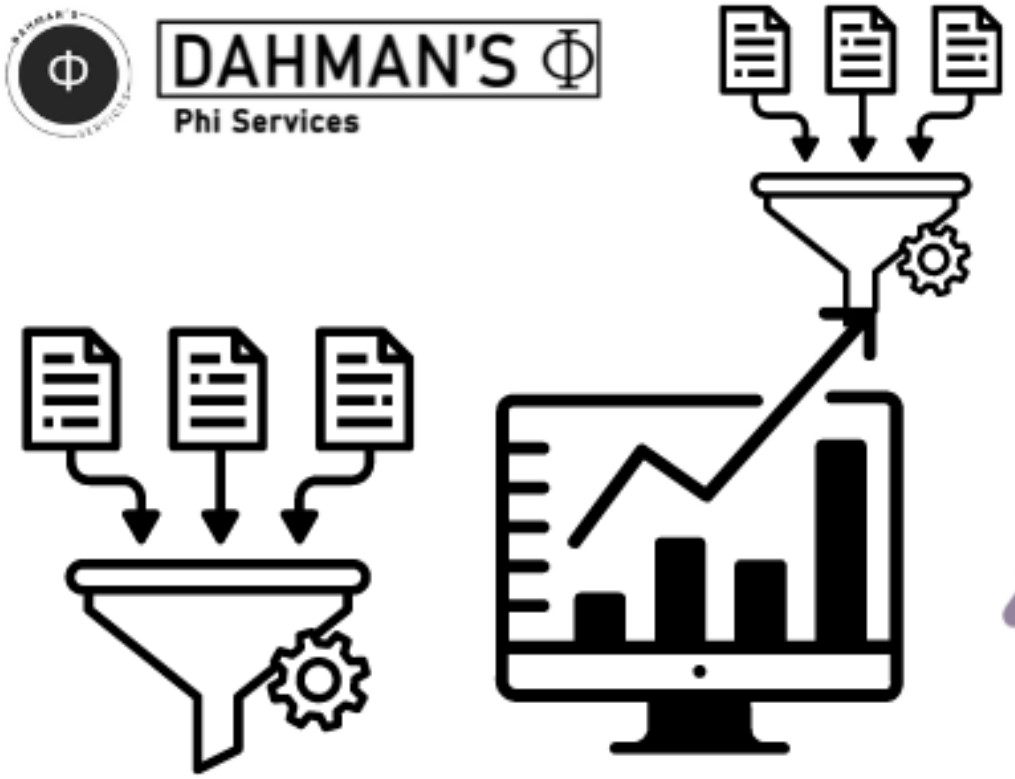
DAHMAN'S Φ
Phi Services

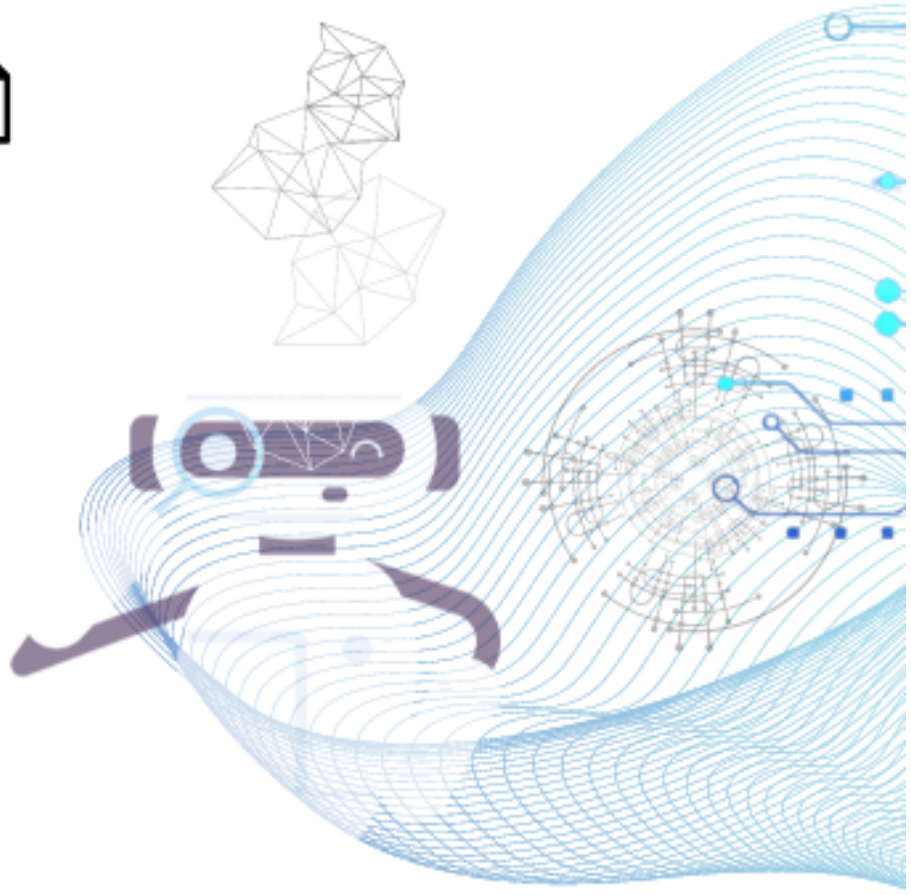


by: Dr. Deniz Dahman's

BOOK TWO

DAHMAN'S Φ
Phi Services





Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

Introduction

- ✓ RESEARCH MAP
- ✓ THREATS TO CONCLUSION VALIDITY
- ✓ STATISTICAL POWER
- ✓ IMPOROVE CONCLUSION VALIDITY
- ✓ ANALYSIS

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

DATA PREPARATION

- ✓ LOGGING THE DATA
- ✓ DATA ACCURACY CONTROL
- ✓ DATABASE STRUCTURE
- ✓ ENTERING DATA TO THE COMPUTER
- ✓ DATA TRANSFORMATION

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

Descriptive Statistics

- ✓ Introduction to EDA
- ✓ Distribution
- ✓ Central Tendency
- ✓ Dispersion
- ✓ Bivariate descriptive
- ✓ Multivariate descriptive

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

Inferential Statistics

- ✓ Introduction
- ✓ Estimating parameters
- ✓ Hypothesis testing

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

STATISTICAL SOFTWARE

- ✓ Introduction
- ✓ Statistical software
- ✓ LAB implementation-

  by JAMOVI

Chapter One

Chapter Two

Chapter Three

Chapter Four

Chapter Five

Chapter Six

Chapter Seven

LAB-SECTION 03

- ✓ Review
- ✓ EDA analysis
- ✓ Inferential Analysis

Chapter One

Chapter Two

Chapter Three

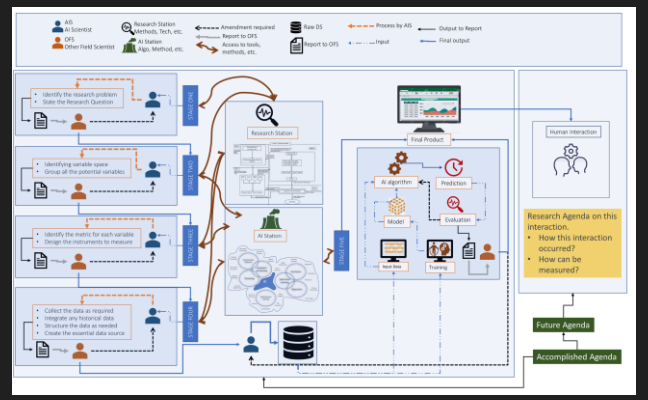
Chapter Four

Chapter Five

Chapter Six

Chapter Seven

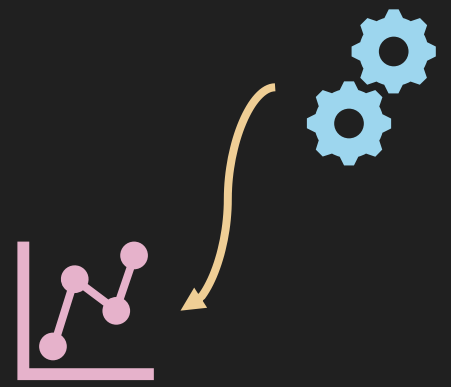
THE MAP



Chapter One

INTRODUCTION

- ✓ RESEARCH MAP
- ✓ THREATS TO CONCLUSION VALIDITY
- ✓ STATISTICAL POWER
- ✓ IMPOROVE CONCLUSION VALIDITY
- ✓ ANALYSIS



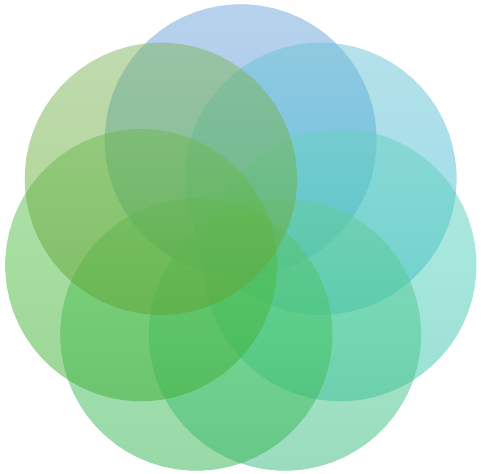
RESEARCH MAP

- ✓ The research from the start to the end

Research question

operationalization

Research structure



Test of measurement

Research design

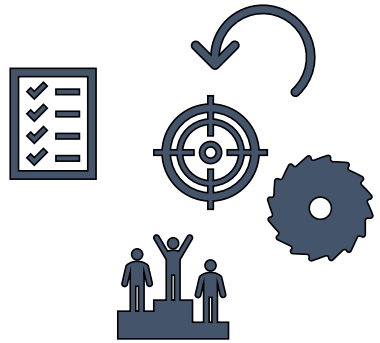
Instrument design

Sampling plan

1	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548
2	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8	64.1	48.8	2548
3	1	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823
4	2	audi 100 ls	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337
5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824
6	2	audi fox	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507
7	1	audi 100ls	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844
8	1	audi 5000	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954
9	1	audi 4000	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086
10	0	audi 5000s (i	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053
11	2	bmw 320i	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395
12	0	bmw 320i	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395
13	0	bmw x1	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710
14	0	bmw x3	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765
15	1	bmw z4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055
16	0	bmw x4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230
17	0	bmw x5	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380
18	0	bmw x3	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	3505

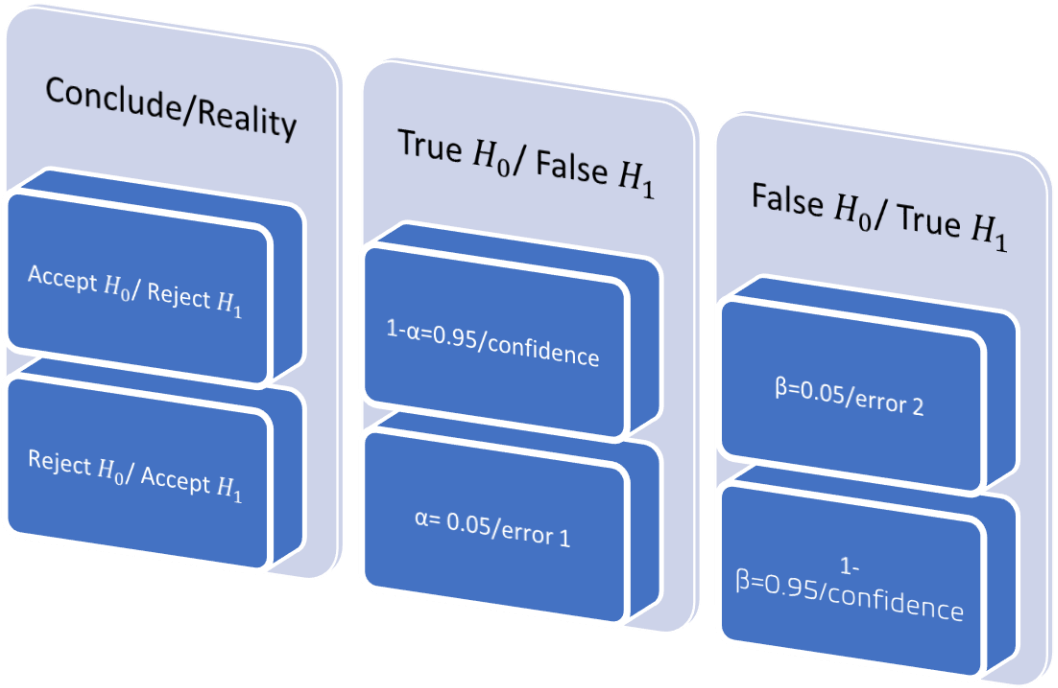
Threats to conclusion validity

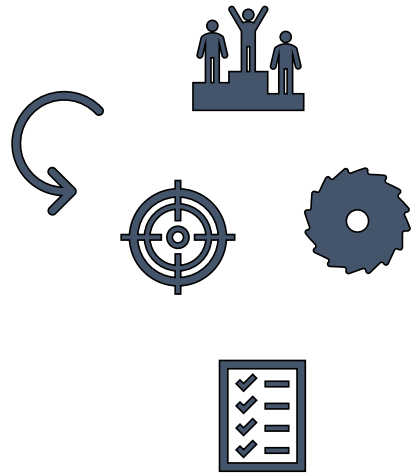
- ✓ Finding no relationship when there is one (or, "missing the needle in the haystack")
- ✓ Finding a relationship when there is not one (or "seeing things that aren't there")
- ✓ Problems that can lead to either conclusion error



Statistical power

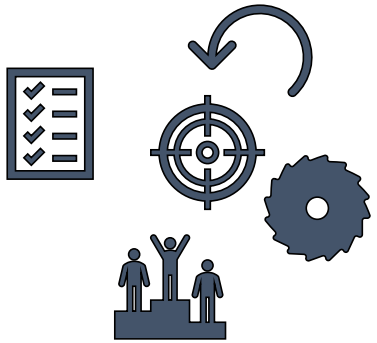
- ✓ The four components (sample size, statistical significance, confidence interval, effect size, alpha level α AND beta level β)
- ✓ The decision matrix:





Improve conclusion validity

- ✓ Good statistical power
- ✓ Good reliability
- ✓ Good implementation



analysis

- ✓ Data preparation
- ✓ Descriptive analysis
- ✓ Inferential analysis

PREDICTION

probability regression

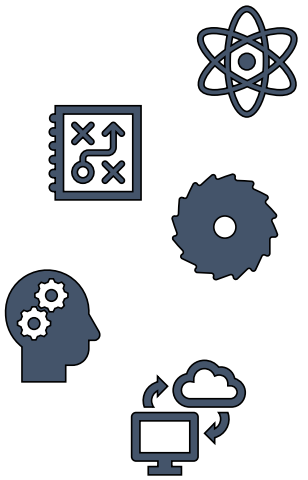
Chapter Two

DATA PREPARATION

- ✓ LOGGING THE DATA
- ✓ DATA ACCURACY CONTROL
- ✓ DATABASE STRUCTURE
- ✓ ENTERING DATA TO THE COMPUTER
- ✓ DATA TRANSFORMATION

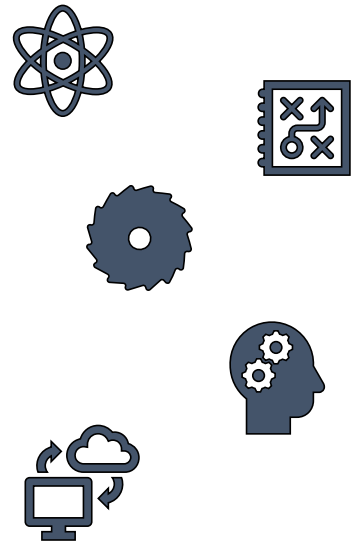
LOGGING THE DATA

- ✓ In any research project you may have data coming from a number of different sources at different times:
  - mail surveys returns
  - coded interview data
  - pretest or posttest data
  - observational data
- ✓ you need to set up a procedure for logging the information and keeping track of it until you are ready to do a comprehensive data analysis
- ✓ You could do this with any standard computerized database program (e.g. MS Access), or standard statistical programs (e.g. SPSS)
- ✓ Most professional researchers will retain such records for at least 5-7 years
- ✓ For important or expensive studies, the original data might be stored in a data archive



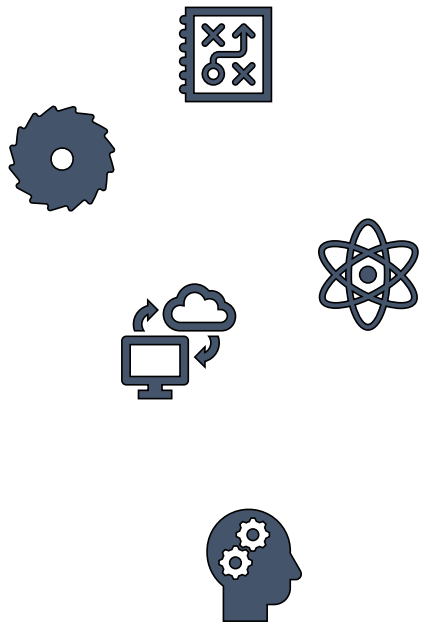
Data Accuracy Control

- ✓ Questions you may consider while you screen the data for accuracy:
  - Are the response readable/legible?
  - Are all important questions answered?
  - Are the responses complete?
  - Is all relevant contextual information included (e.g. data, time, place, researcher)?
- ✓ Benefits for such approach:
  - doing this right away will allow you to go back to the sample to clarify any problems or errors
  - As most research stress on the fact that quality of measurement is major issue, Assuring that the data collection process does not contribute inaccuracies will help assure the overall quality of subsequent analyses



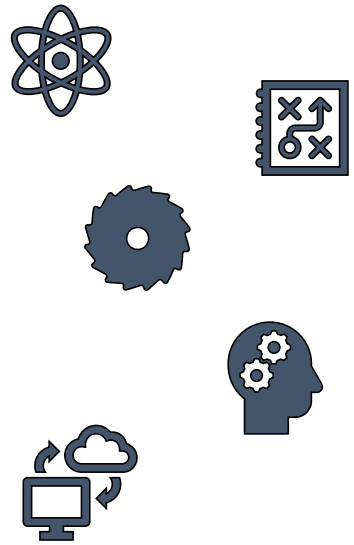
Database Structure

- ✓ manner in which you intend to store the data for the study
- ✓ in large complex studies, you might have one structure for logging data and another for storing it
- ✓ you should generate a printed codebook that describes the data and indicates where and how it can be accessed
- ✓ codebook should include the following items for each variable:
  - variable name
  - variable description
  - variable format (number, data, text)
  - instrument/method of collection
  - date collected
  - respondent or group
  - variable location (in database)
  - Notes
- ✓ codebook together with the database, provide comprehensive documentation for further analysis you need no additional information



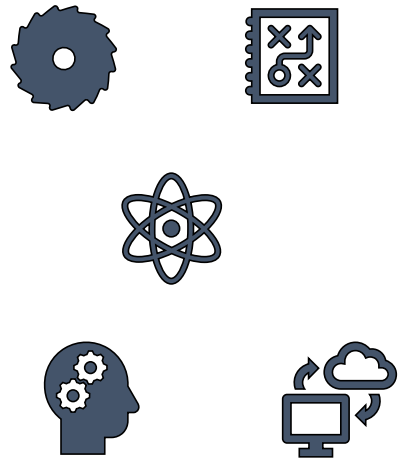
Entering Data to Computer

- ✓ Simply to enter your data to the computer, available approaches to control the accuracy could be
- ✓ Double entry approach: enter the data once, and then second time, automated app will discover if any discrepancy
- ✓ Procedural entry approach: enter the data once, and then set up procedure for checking the data for accuracy, e.g. available software controls that by summary of data, e.g. age is 606, or maybe 7 is entered when the expected value is only 1 to 5



Data transformation

- ✓ it is almost always necessary to transform the raw data into variables that are usable in the analyses
- ✓ Some of the more common are:
  - Item reversal
  - Scale total
  - Categories



Chapter Three

Descriptive Analysis

- ✓ Introduction to EDA
- ✓ Distribution
- ✓ Central Tendency
- ✓ Dispersion
- ✓ Bivariate descriptive
- ✓ Multivariate descriptive

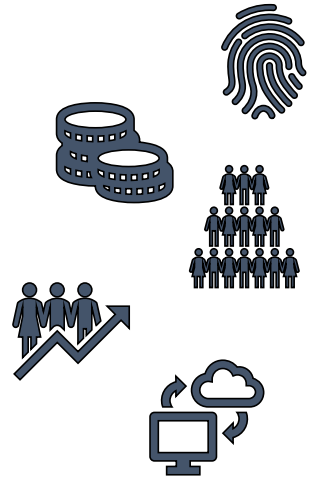
Introduction to EDA

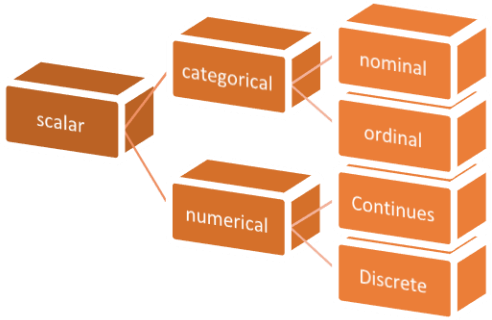
DESCRIPTIVE=STATISTICS

PROBABILITY HAS NOTHING TO DO IN HERE

- ✓ To answer research question:
  - The variable
- ✓ To quantify the variable:
  - The value (scalar)/ outcome
- ✓ many subjects of ONE variable on that scalar:
  - The univariate (vector): to analyze it:

    (1) distribution, (2) central tendency, (3) dispersion
- ✓ To analyze two variable scalars/vectors:
  - The bivariate (column)
- ✓ To analyze more than two variables:
  - The multivariate (matrix)



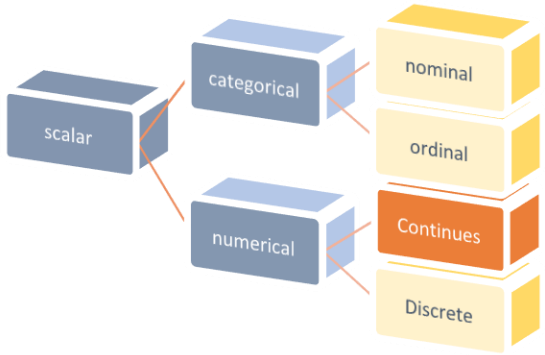


Distribution

- ✓ The distribution is a summary of the frequency of individual values or ranges of values for a variable
- ✓ Format of distribution:
  - Table distribution (frequency distribution)
  - Graph distribution (counting distribution)
- ✓ Explore distribution in terms of the scalar:

MASS

DENSITY

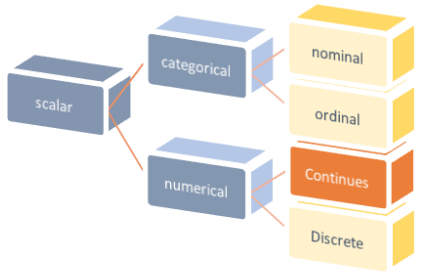


Distribution

- ✓ Illustration of the (variable outcome in form of scalar or value where it's mass)
- ✓ illustrate the number of fruits in the fridge that might only have (apple, banana, orange)
- ✓ Common mass distribution graph:
  - Bar
  - Pie
  - customized
- ✓ Distribution Table

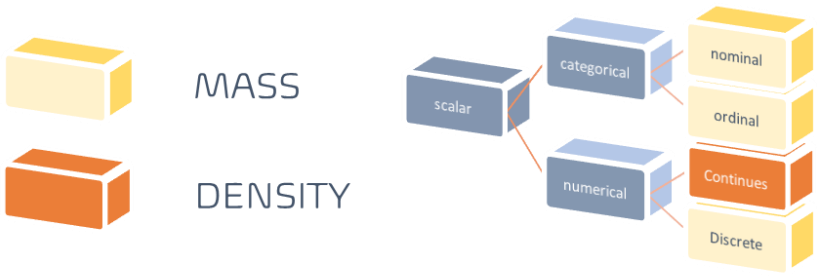
MASS

DENSITY



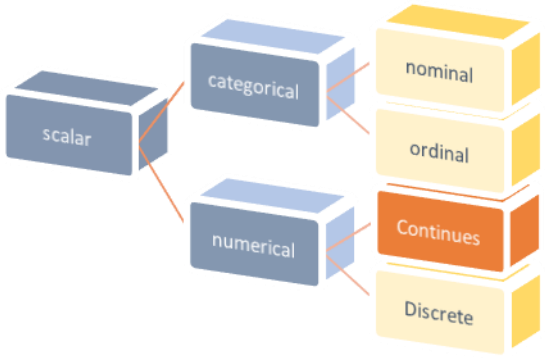
Distribution

- ✓ Illustration of the (variable outcome in form of scalar or value when it's density)
- ✓ illustrate the element weight in gram
- ✓ Common mass distribution graph:
  - Histogram
  - customized
- ✓ Distribution table



Central Tendency

- ✓ Illustration of the (average/expectation)
- ✓ Illustration on the center point
- ✓ Common calculations on the center tendency:
  - Mean
  - Median
  - Mood
  - Proportion
- ✓ Story of outliers



Central Tendency

- ✓ Illustration of the (mean)
- ✓ Illustration on the difference of symbols (μ; $\bar{x}$)
- ✓ To find the mean, simply add up all response values and divide the sum by the total number of responses. The total number of responses or observations is called N.
- ✓ Graphical illustration

Central Tendency

- ✓ Illustration of the (median)
- ✓ The median is the value that's exactly in the middle
- ✓ To find the median, order each response value from the smallest to the biggest. Then, the median is the number in the middle. If there are two numbers in the middle, find their mean.
- ✓ Graphical illustration

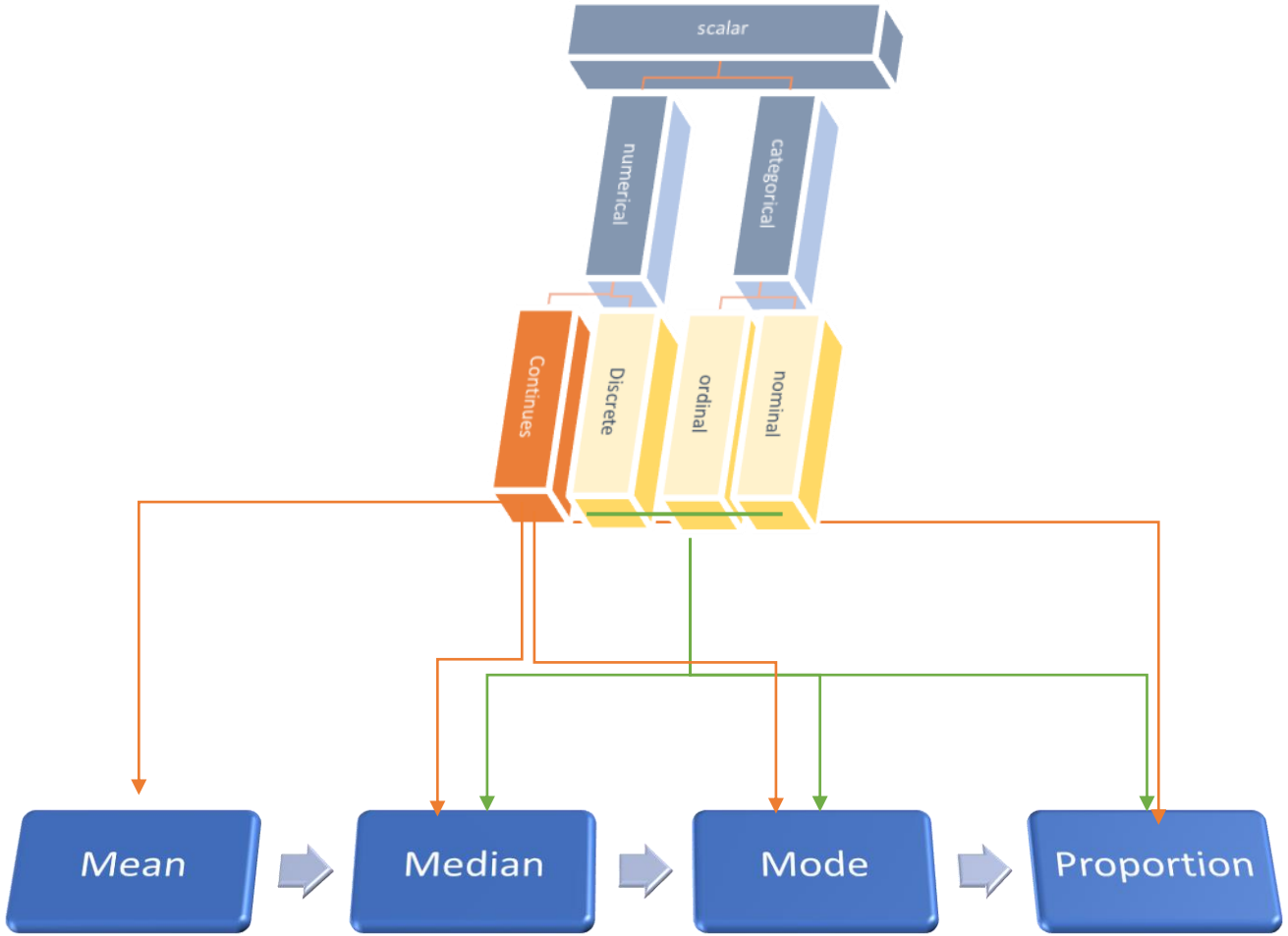
Central Tendency

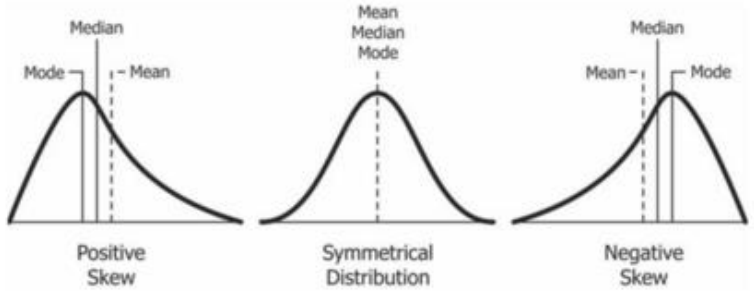
- ✓ Illustration of the (mode)
- ✓ the most popular or most frequent response value
- ✓ To find the mode, order your data set from lowest to highest and find the response that occurs most frequently.
- ✓ Graphical illustration

Central Tendency

- ✓ Illustration of the (proportion)
- ✓ A population proportion is a fraction of the population that has a certain characteristic
- ✓ Its relation to ordinal data.
- ✓ Graphical illustration

Central Tendency





Dispersion

- ✓ Illustration of the (dispersion)
- ✓ Variability describes how far apart data points lie from each other and from the center of a distribution
- ✓ Variability is also referred to as spread, scatter or dispersion
- ✓ Tools:
  - Range
  - Interquartile
  - Variance
  - Standard deviation

Dispersion

- ✓ Illustration of the (range)
- ✓ The range tells you the spread of your data from the lowest to the highest value
- ✓ To find the range, simply subtract the lowest value from the highest value in the data set.
- ✓ Because only 2 numbers are used, the range is influenced by outliers and doesn't give you any information about the distribution of values.
- ✓ It's best used in combination with other measures

Dispersion

- ✓ Illustration of the (interquartile range)
- ✓ The interquartile range gives you the spread of the middle of your distribution
- ✓ The interquartile range is the third quartile (Q3) minus the first quartile (Q1). This gives us the range of the middle half of a data set
- ✓ the interquartile range uses only 2 values in its calculation. But the IQR is less affected by outliers.
- ✓ The Five-number summary:
  - Lowest value
  - Q1: 25th percentile
  - Q2: the median
  - Q3: 75th percentile
  - Highest value

Dispersion

- ✓ Illustration of the (variance)
- ✓ A deviation from the mean is how far a score lies from the mean
- ✓ Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean
- ✓ it's harder to interpret the variance number intuitively, it's important to calculate variance for comparing different data sets in statistical tests like ANOVAs

Dispersion

- ✓ Illustration of the (standard deviation)
- ✓ The standard deviation is the average amount of variability in your dataset
- ✓ It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is
- ✓ The difference between biased and conservative estimates of standard deviation gets much smaller when you have a large sample size

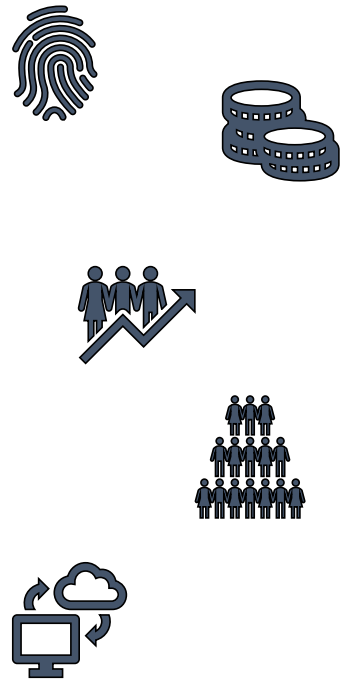
Bivariate descriptive

- ✓ Illustration of the (bivariate descriptive)
- ✓ study the relationship that exists between two variables

  Correlation Coefficients –The coefficient lets you know if the data in question are related. When the correlation coefficient is zero, the variables are not related. If the correlation coefficient is a positive or a negative 1, then this means that the variables are perfectly correlated
- ✓ Tools:
  - Scatter plots – This gives an idea of the patterns that can be formed using the two variables
  - contingency table displays frequencies for combinations of two categorical variables.

Multivariate descriptive

- ✓ Illustration of the (multivariate descriptive)
- ✓ involves analyzing relationships between more than two variables
- ✓ Main Tools:
  - All what you know about uni-bi variate
  - covariance and correlation matrices.



Chapter Four

Inferential Analysis

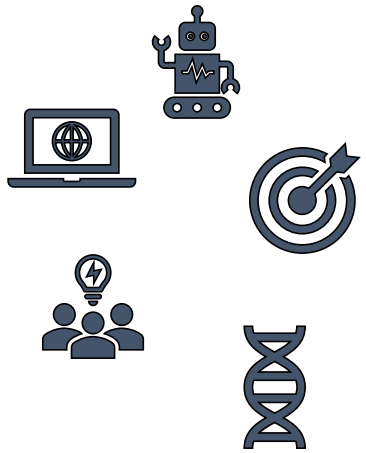
- ✓ Introduction
- ✓ Estimating parameters
- ✓ Hypothesis testing

Introduction

- ✓ Illustration of the (inferential descriptive)
- ✓ While descriptive statistics can only summarize a sample's characteristics, inferential statistics use your sample to make reasonable guesses about the larger population
- ✓ With inferential statistics, it's important to use random and unbiased sampling methods

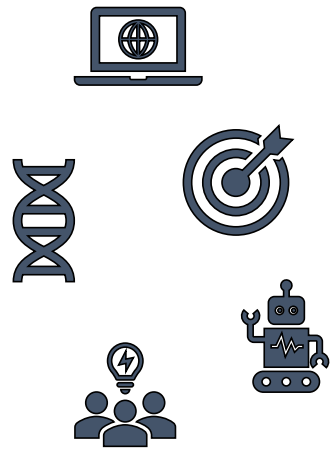
  characteristics of samples and populations ($\mu \hat{y} s \sigma p$):

  - A statistic is a measure that describes the sample (e.g., sample mean).
  - A parameter is a measure that describes the whole population (e.g., population mean)



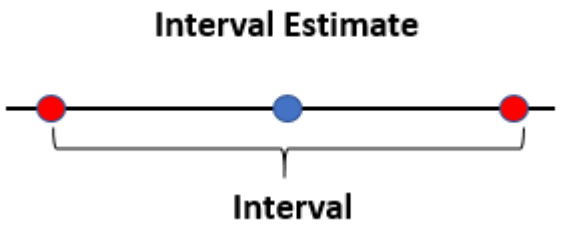
Estimating parameters

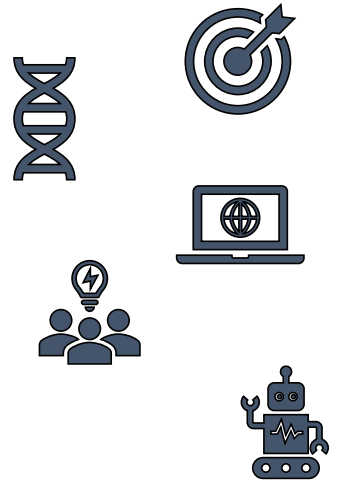
- ✓ This means taking a statistic from your sample data (*nŷsp*) and using it to say something about a population parameter (*NμσP*)
- ✓ Steps:
  - Select random and representative sample
  - Collect info from the sample
  - Calculate the sample statistics of interest
  - Assign value to the population parameter



Estimating parameters

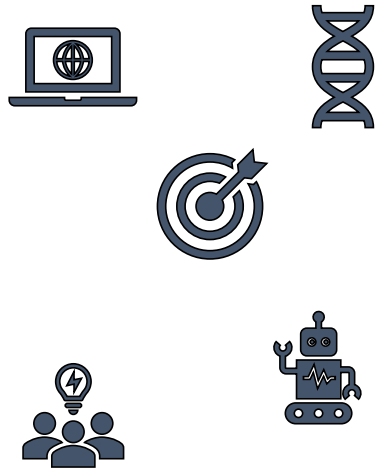
- ✓ A confidence interval estimate is a range of values constructed from sample data so that the population parameter is likely to occur within the range at a specified probability.
- ✓ factors that determine the width of a confidence level:
  - Sample size
  - Population variability
  - Desired level of confidence
- ✓ Confidence Interval





Hypothesis testing

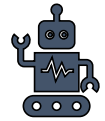
- ✓ The main purpose of statistics is to test a hypothesis
- ✓ A hypothesis is an educated guess about something in the world around you.
- ✓ It should be testable, either by experiment or observation. For example:
  - A new medicine you think might work.
  - A way of teaching you think might be better.
  - A possible location of new species.
  - A fairer way to administer standardized tests.
- ✓ Visual illustration on the right test as if it is the right tool



Hypothesis testing

- ✓ Type of the variable based on the idea of (numerical or categorical)
- ✓ Number of samples
  - One, you test against the hypothesized value
  - One sample and each subject is measured on different variable
  - Two or more, and you compare to each other
- ✓ Type of analysis- purpose
  - Test against the hypothesized values/points or i.e. the claim
  - Comparison of two statistics
  - Relationship
- ✓ Distribution
  - Normal
  - Non-normal
  - Dichotomous



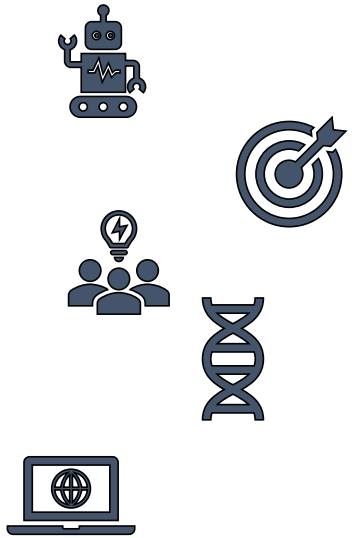


Outcome	continues										discrete		
# samples	One – one Var		One – Two Var			Not One					One – one Var	One – Two Var	
						Two		> two					
purpose	Hypothesized value		Comparison		Relationship	Comparison					Hypothesized value	Comparison	Relationship
	mean	variance	mean	variance		mean	variance	mean			Proportion		
								One factor	Two factor	< two factor			
										variance			
test	Test of mean T-test	Chi square for variance	Paired sample t-test	F-test	• Pearson's test • Spearman's test	Independent sample t-test	• F-test • Bartlett test • Leven test	ANOVA-One way	• ANOVA-Two way • ANCOVA	• MANOVA • MANCOVA	• F-test • Bartlett test • Leven test	test for proportion	Z-test for proportion	Chi square for independence

	Nominal		Ordinal	Continues		
One Sample	Hypothesized value			Hypothesized value		Hypothesized value
	Proportion		*Median*	*Mean*		*Variance*
	Binomial Test test for proportion One Sample Chi-Square Test		One-sample median test	T-test one sample		Chi square for variance
Two sample						
Independent	*Proportion*		*Median*	*Mean*		*Variance*
	z-test for independent proportions		Median test Mann-Whitney U Test	t-test for homogeneous variances t test for heterogeneous variances		F-test
Dependent	*Proportion*		*Median*	*Mean*		*Variance*
	Comparison	**Relationship**		**Comparison**	**Relationship**	
	Paired Samples Z-Test McNemar-Test	Chi square test	Wilcoxon-Test	Paired t test	Pearson's test Spearman's test	F-test
> Two samples						
Independent	*Proportion*		*Median*	*Mean*		*Variance*
	Chi-Square Test		H-Test Kruskal & Wallis	One factor	ANOVA one way/wo repeat	Bartlett test Leven test
				Two factors	ANOVA two way/wo repeat ANCOVA/wo repeat	
				> Two factors	MANOVA/ wo repeat MNCOVA/ wo repeat	
Dependent	Cochran-Test		Friedman-Test	One factor	ANOVA one way/w repeat	Bartlett test Leven test
				Two factors	ANOVA two way/w repeat ANCOVA/wo repeat	
				> Two factors	MANOVA/ w repeat MNCOVA/ w repeat	

Hypothesis testing

	Comparison				Relationship
	2 dataset		> 2 dataset		TWO VARIABLES
	Paired	Unpaired	Paired	Unpaired	
Normal distribution	P/ t test	Up/ t test	Repeated measures of ANOVA	One way ANOVA	Person correlation
Non-Normal distribution	Wilcoxon signed rank	• Wilcoxon rank sum • Mann Whitney U test	Friedman test	Kruskal Wallis test	Spearman's rank correlation
Dichotomous distribution	McNemar's test	• Chi square test • Fisher exact test	Cochran's Q test	Chi square test	Contingency coefficient



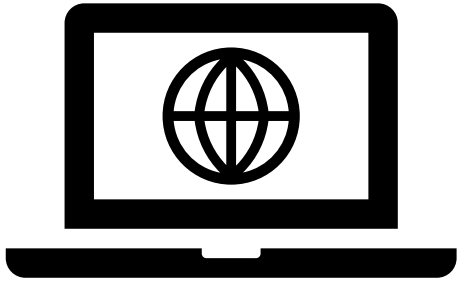
Chapter Five

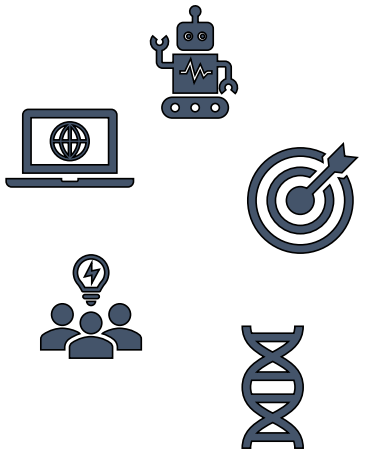
STATISTICAL SOFTWARE

- ✓ Introduction
- ✓ Statistical software
- ✓ LAB implementation

Introduction

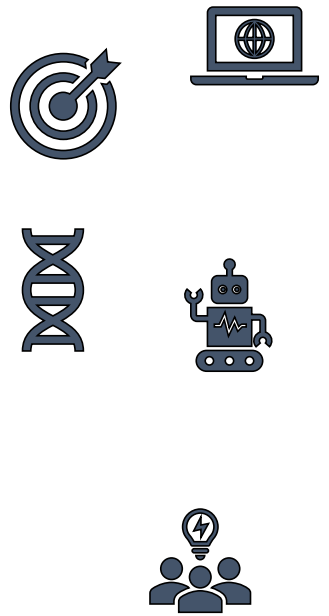
- ✓ Illustration of the (statistical software)
- ✓ specialized computer programs for analysis in statistics and econometrics
- ✓ Code or software ?





Statistical Software

- ✓ Illustration of the (available options)
- ✓ Open source
  - ADaMSoft
  - Jamovi
  - R
- ✓ Public domain
  - CSPro
  - Dataplot
- ✓ Freeware
  - MaxStat Lite
  - MINUIT
- ✓ Proprietary
  - SAS (software)
  - SPSS Statistics
  - S-PLUS



Lab implementation

- ✓ Illustration of the (Jamovi)
- ✓ Software illustration
- ✓ BUG DATASET ANALYSIS

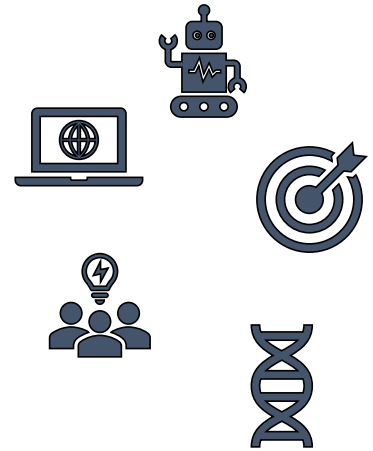
Chapter Six

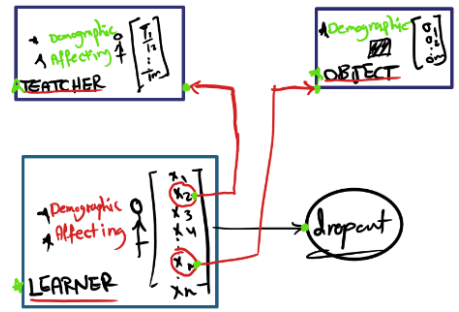
LAB-SECTION_03

- ✓ Review
- ✓ EDA analysis
- ✓ Inferential Analysis

Review

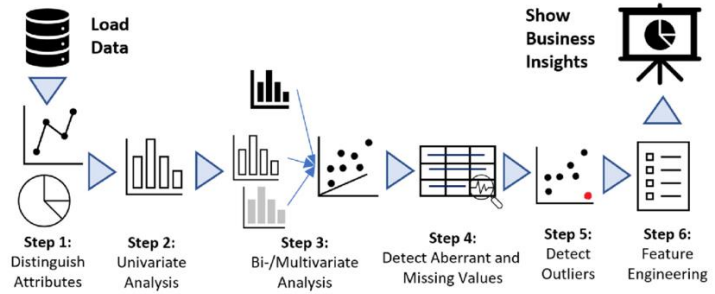
- ✓ Research Question:
  - do the demographic & the affecting variables affect the decision of a life long learner to drop out a life-long training course?
  - can those variables, collectively, subgroup or individually be predictive to the decision of dropping out the course?
- ✓ Research Hypothesis:
  - H_0: the demographic & affecting variables, collectively or subgroup have NO affect on the decision of a life-long learner to drop-out a life-long training course.
  - H_1: the demographic & affecting variables, collectively or subgroup have affect on the decision of a life-long learner to drop-out a life-long training course

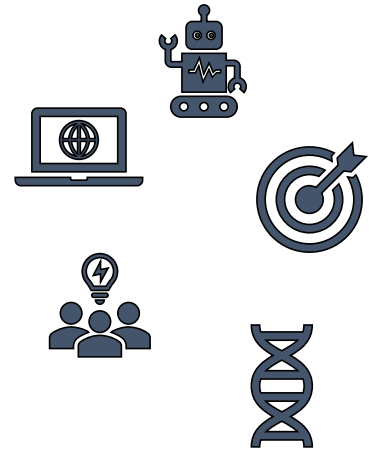




EDA Analysis

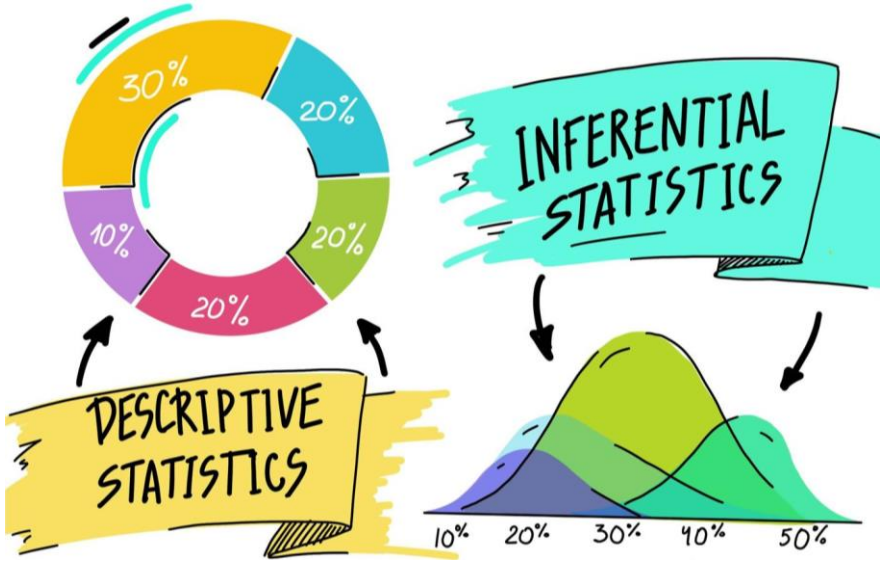
- ✓ Organize
- ✓ Clean data
- ✓ Missing
- ✓ Descriptive model

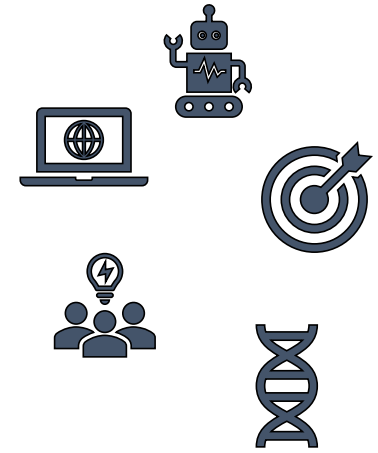




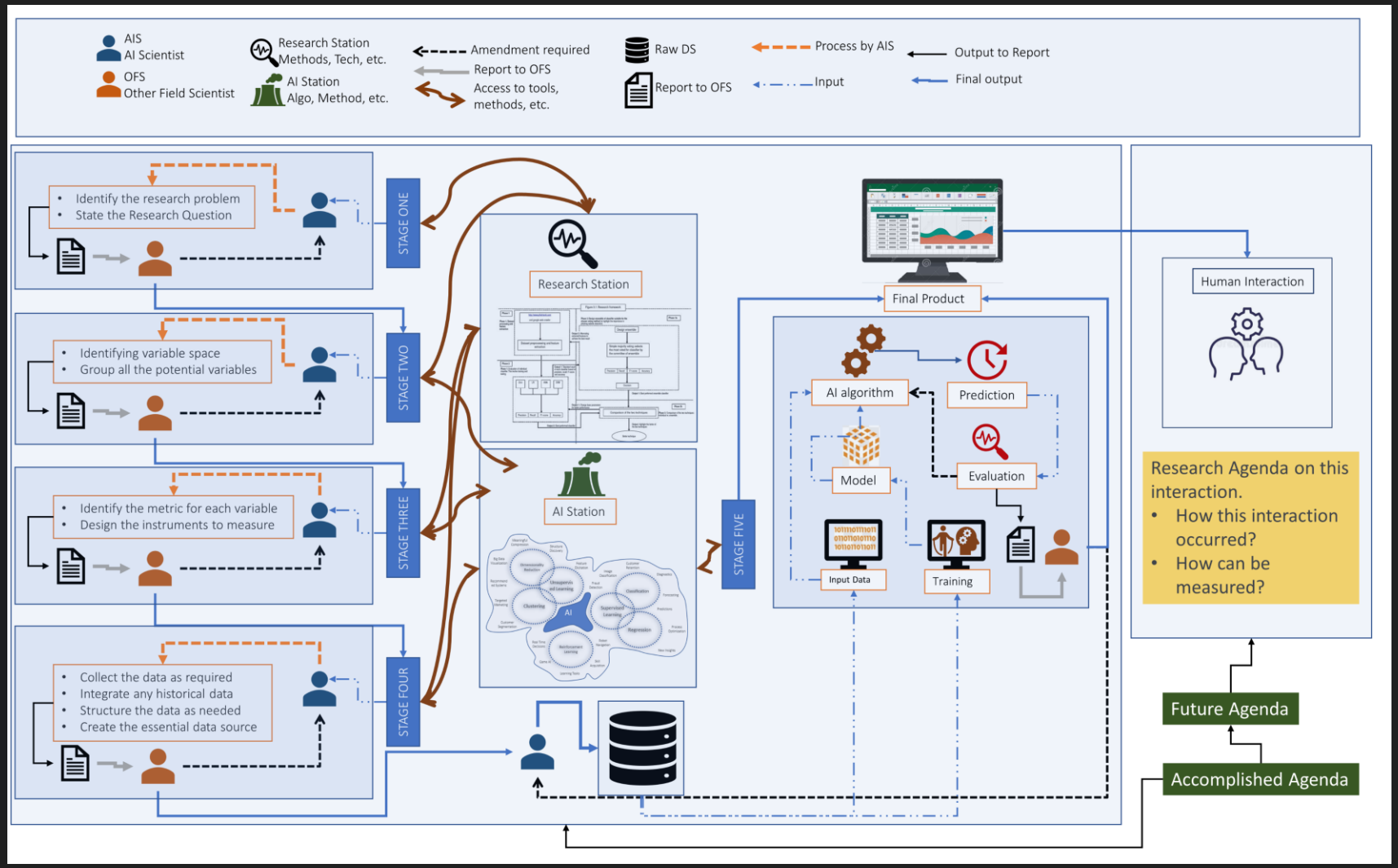
Inferential Analysis

- ✓ Explore the sample
- ✓ Determine the estimates
- ✓ Test the hypothesis





THE MAP



AISS V2.1.0