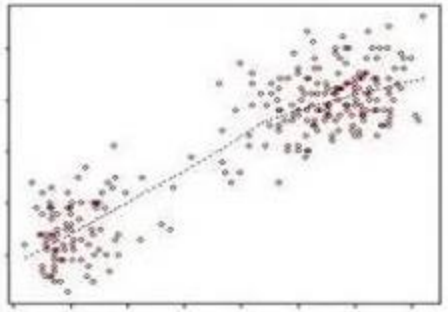
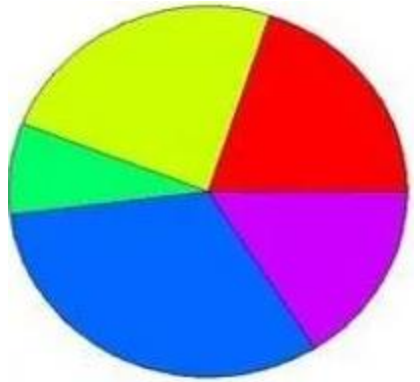
VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

Statistics using Python: Course Outline

Jyostna Devi Bodapati (PhD)
Asst. Prof, CSE, VFSTR





COURSE: Statistics using Python (SUP)

Syllabus: UNIT- I

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

- Why Statistics?
- Python Packages for Statistics
- Review of Python Programming
- Pandas: Data Structures for Statistics

- Data Input: Input from Text Files, Visual Inspection, Reading ASCII-Data into Python, Input from MS Excel

- Data types: Categorical, Numerical.

Syllabus: UNIT- II

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

Displaying Statistical Datasets:

- **Univariate Data:** Scatter Plots, Histograms, Kernel-Density-Estimation (KDE) Plots, Cumulative Frequencies, Error-Bars, Box Plots, Grouped Bar Charts, Pie Charts.
- **Bivariate and Multivariate Plots:** Bivariate Scatter Plots
- **3-D Plots**

Syllabus: UNIT- III

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

- Populations and Samples
- **Distribution Center:** Mean, Median, Mode, Geometric Mean
- **Quantifying Variability:** Range, Percentiles, Standard Deviation and Variance.
- **Discrete Distributions-** Bernoulli Distribution, Binomial Distribution, Poisson Distribution,

Syllabus: UNIT- IV

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

- **Normal Distribution-** Examples of Normal Distributions, Central Limit Theorem
- Continuous **Distributions Derived from the Normal Distribution:** t-Distribution, Chi-Square Distribution, F-Distribution.
- **Hypothesis Tests:** Typical Analysis Procedure: Data Screening and Outliers, Normality Check, Hypothesis Concept, Errors, **p-Value**, and Sample Size-Generalization and Applications, **The Interpretation of the p-Value,**
- Types of Error, Sensitivity and Specificity.

Syllabus: UNIT- IV

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

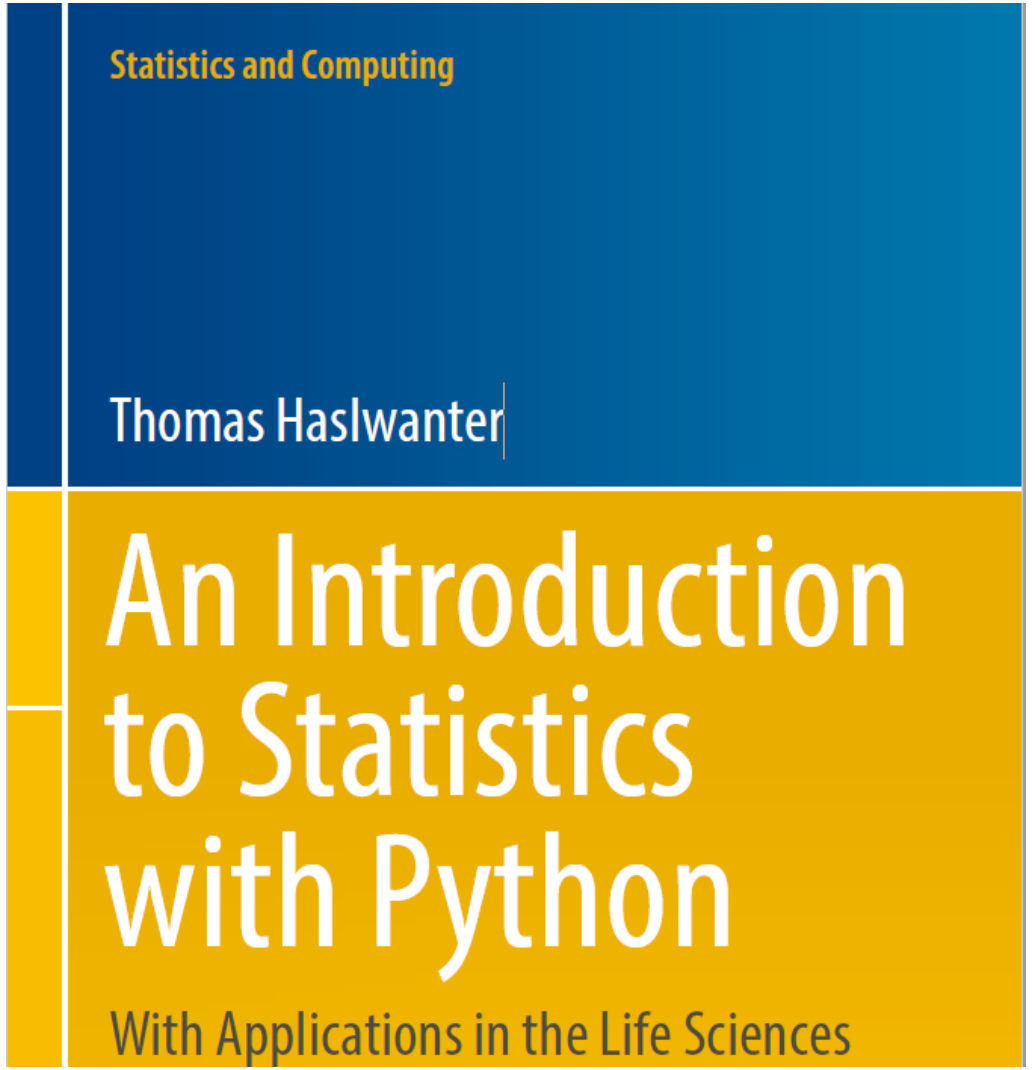
- Analysis of Variance (ANOVA)-One-Way ANOVA, Two-Way ANOVA, One-Way Chi-Square Test, Chi-Square Contingency Test
- Linear Regression Models-Linear Correlation-Correlation Coefficient, Rank Correlation, General Linear Regression Model, Coefficient of Determination, Linear Regression Analysis with Python.

Book for Reference

"An Introduction to Statistics with Python With Applications in the Life Sciences", Thomas Haslwanter - Springer- ISSN 1431-8784 - ISBN 978-3-319-28315-9, Springer International Publishing, Switzerland 2016.

Book for Reference

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956



Evaluation

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

Theory Course:

Teaching:

4 Lecture Hours/ week

Recommend: Practice the coding exercises

Evaluation:

Internal Marks: 40M

Week Tests + Mid Exams

External Marks: 60M

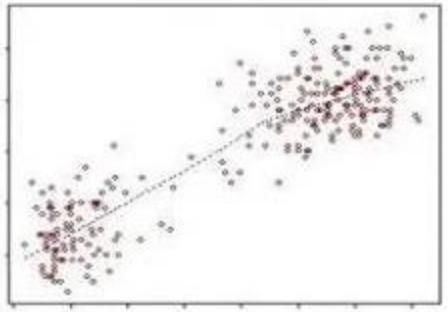
End Semester Exam


SUMMARY

Thank You

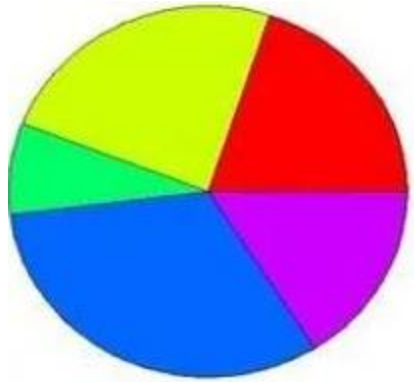
VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

Introduction to Statistics



Jyostna Devi Bodapati (PhD)

Asst. Prof, CSE, VFSTR



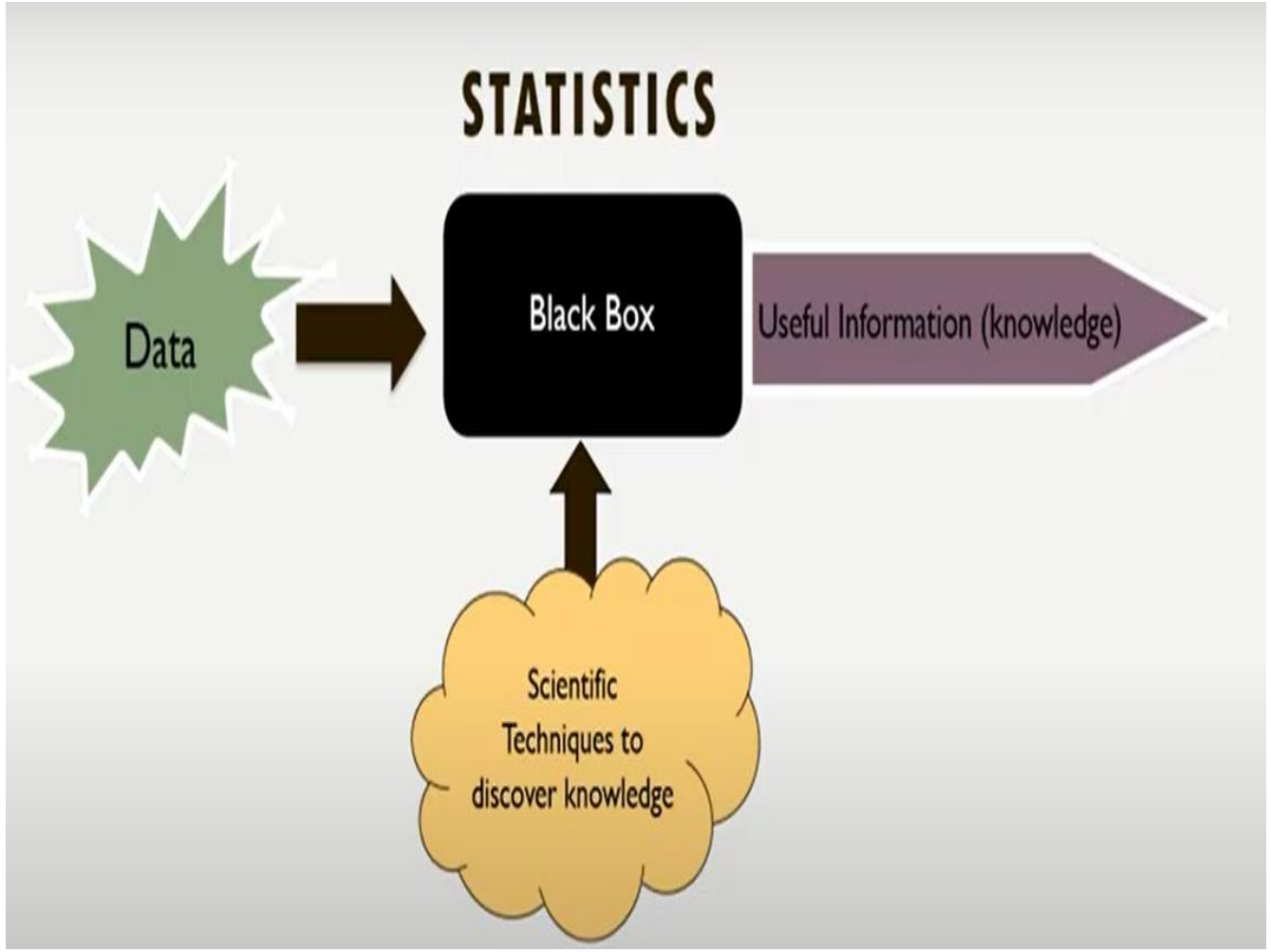
COURSE: Statistics using Python (SUP)

What is Statistics?

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

- Statistics is the science of collecting, organizing, summarizing, analyzing, and making inferences from data
- Statistics are the sets of mathematical equations that we use to analyze the available data. It Provides information
- Useful in taking Decisions
- The field of statistics is the science of learning from data

What is Statistics?



Why Statistics?

Statistical knowledge helps to:

- **Use proper methods to collect the data,**
- **Employ the correct analyses and**
- **Effectively present the results.**
- **Make quantitative statements about estimated parameters.**
- **Make future predictions based on the data.**

Applications of Statistics?

Statistics helps to:

- Weather forecasting
- Online shopping
- Politics
- Insurance
- Stock market
- Sports
- Medical
- Agriculture
- Emergency Preparedness
- Genetics
- Consumer Goods

Use of tools for statistics

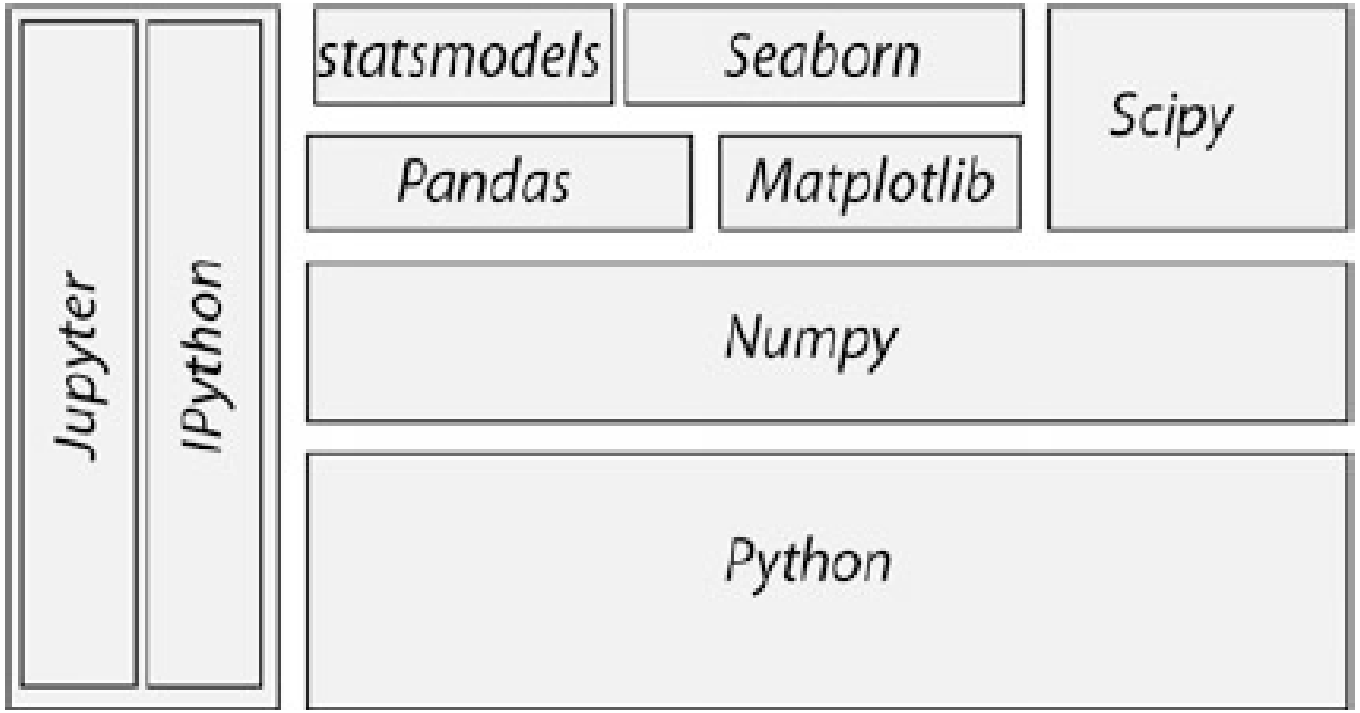
- Excel
- R
- Python
- BI

Note: Python can be used across domains

Python Libraries for Statistics

Python Libraries for Statistics

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956



The structure of the *Python* packages

Python Distributions

- Popular Python distributions are:
  - WinPython
  - ActivePython
  - Cpython
  - Anaconda

Python Distributions

- **WinPython:**
  - Recommended for Windows users.
  - Free and customizable.
  - Latest version is 3.8.7.

- **Anaconda:**
  - Recommended for Windows, Mac, and Linux.
  - Can be used to simultaneously install Python 2.x and 3.x
  - The latest Anaconda version is 5.3.0.
  - Anaconda is free for educational purposes.

Python packages

- Ipython
- numpy
- Scipy
- Matplotlib
- Pandas
- patsy
- Statsmodels
- Seaborn
- xlrd
- PyMC
- scikit-learn
- scipy
- lifelines
- rpy2

Python Libraries for Statistics

- Many popular Python toolboxes/libraries:
  - NumPy
  - SciPy
  - Pandas
  - Statsmodels

- Visualization libraries
  - matplotlib
  - Seaborn

  - and many more …

Python packages

- **ipython:**
  - An upgraded Python read-eval-print loop (REPL) for interactive work.
- **Numpy:**
  - Supports working with vectors and arrays.
- **Pandas:**
  - Data manipulation
- **Matplotlib:**
  - The de-facto standard module for plotting and visualization.
- **Seaborn:**
  - For visualization of statistical data.

Python packages

- **Patsy:**
  - For working with statistical formulas.
- **Statsmodels:**
  - For statistical modeling and advanced analysis.
- **Scipy:**
  - All the essential scientific algorithms, including those for basic statistics.
- **PyMC:**
  - For Bayesian statistics, including Markov chain Monte Carlo simulations.
- **scikit-learn:**
  - For machine learning.

Python packages

- **scikits.bootstrap:**
  - Provides bootstrap confidence interval algorithms for scipy
- **Lifelines:**
  - Survival analysis in Python.
- **rpy2:**
  - Provides a wrapper for R-functions in Python.
- **Xlrd:**
  - For reading and writing MS Excel files

Installation of Python Libraries

PyPI (The Python Package Index)

- PYPI is a repository of software for the *Python* programming language
- Currently with more than 80,000 packages.
- Packages from *PyPI* can be installed easily, from the Windows command shell (cmd) or the Linux terminal, with:
  - $pip install [_package_]
- To update a package, use:
  - $pip install [_package_] –U

➢ To list all the installed packages

- $ pip list

Install Pandas

- **Pip Installer:**

  $ pip install pandas

- **Conda Installer:**

  $ conda install pandas

- **Jupyter Notebook:**

  !pip install pandas

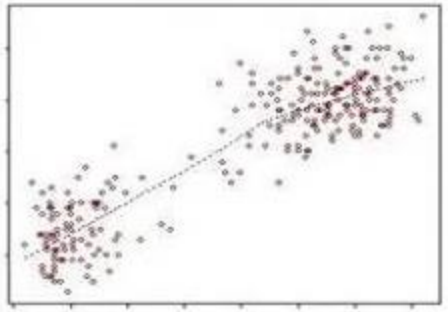

SUMMARY

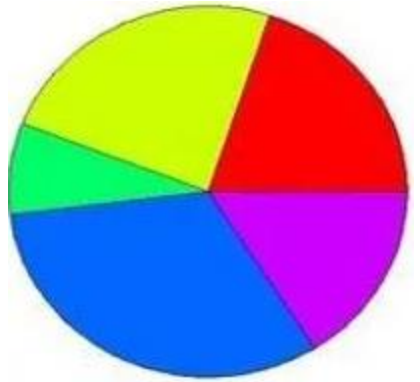
Thank You

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

Data Types

Jyostna Devi Bodapati (PhD)
Asst. Prof, CSE, VFSTR



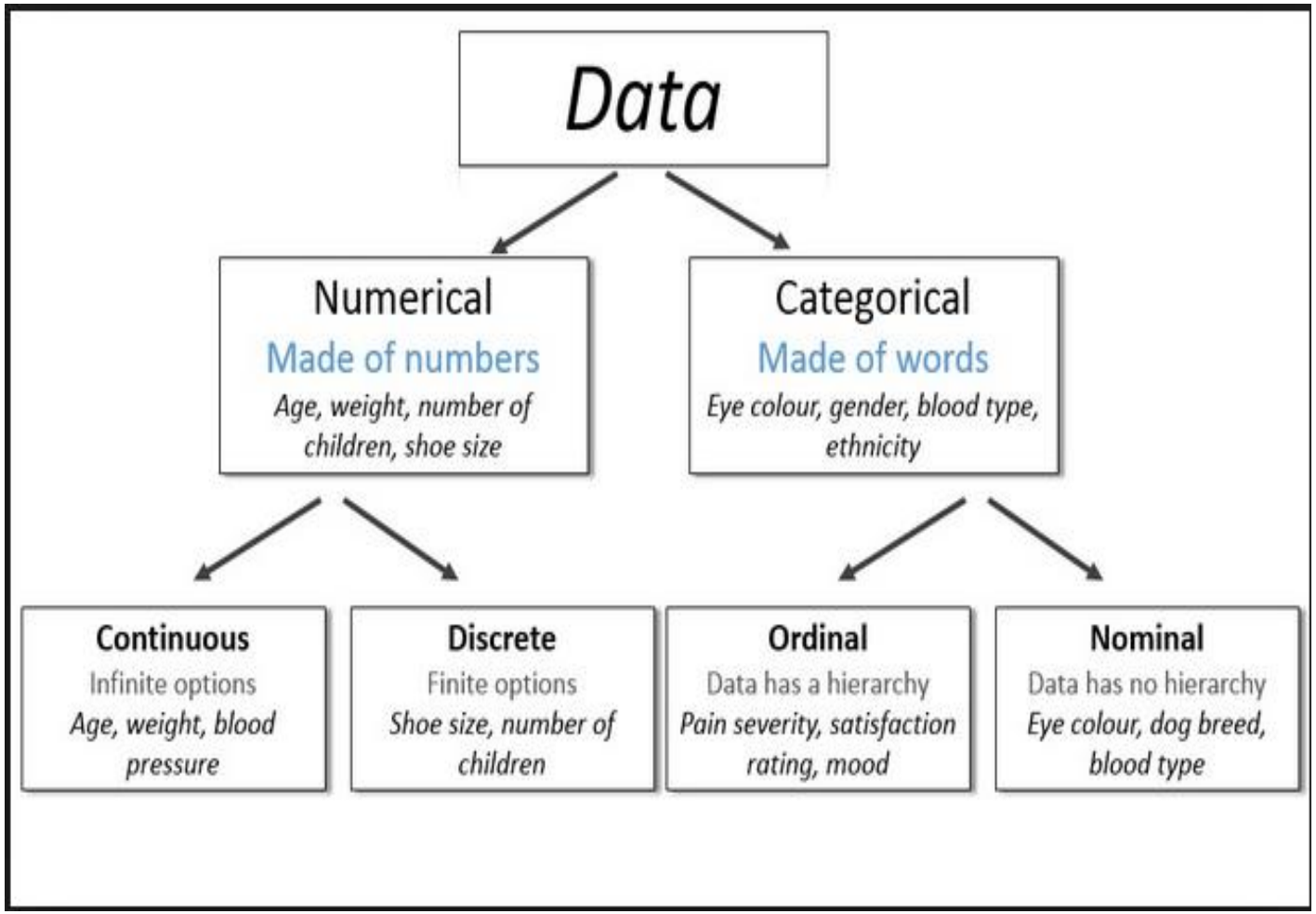


COURSE: Statistics using Python (SUP)

Data Types

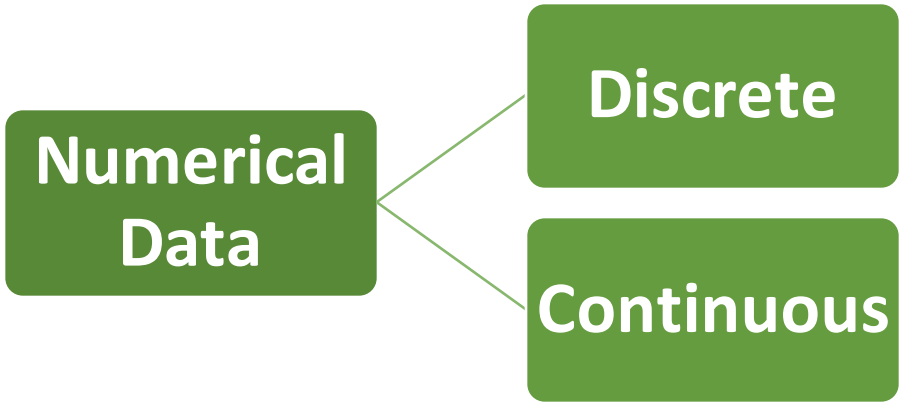
- Data refers to the collected raw facts.
- This data could be of any type
- Data is often used to prove or disprove a hypothesis or scientific guess, during an experiment.

Types of data



Numerical Data

- The type of data which can be measured
- Also known as quantitative data
- Ex: person's height, weight, IQ, or blood pressure; number of shares, teeth a dog, pages in a book



Discrete Data

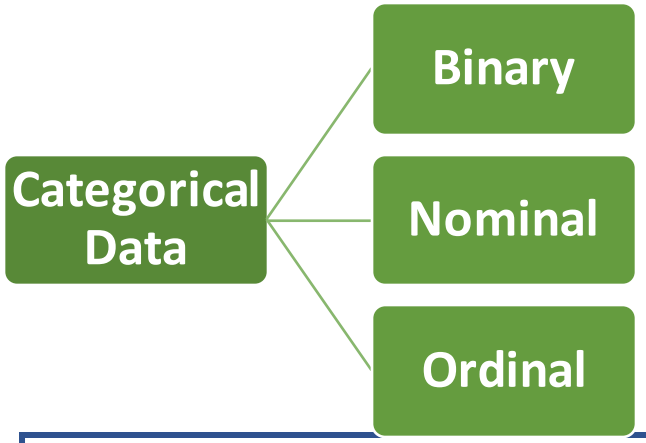
- Discrete data has distinct set of values, which are countable and belonging to whole numbers set (0 1 2 3 ….)
- It cannot take the values of a fractions
- Examples:
  - Number of students
  - Number of days rained in a year
  - Number of children in the family

Continuous Data

- Continuous data refers to any values within an interval.
- Cam be any value with in the range
- Value can be fractional / real
- Examples:
  - Height of students,
  - Rainfall in an year
  - Time
  - Temperature

Categorical Data

- The values that describe a quality or characteristic of data like what type or what category.
- They fall into mutually exclusive (in one category or another) and exhaustive (include all possible options) categories.
- These are qualitative variables (non numeric values)



Boolean data

- Boolean data refers to the data that can take one of the two possible values.

- Examples:
  - Gender: female/male
  - Result: Pass/Fail
  - Married: True/False
  - Taste: Good/bad

Nominal Data

- Nominal data refers to the data that can take one of the possible values from the given set.
- No ordering among the data.
-
- **Examples:**
  - Color of the Shirt: red, blue, yellow
  - Type of fruit: Apple, banana,
  - Marital status: Unmarried, married, divorced/separated, widowed

Ordinal Data

- Ordinal data refers to the data that can take one of the possible values from the set.
- Ordering among the data exists.

- **Examples:**
  - Rank
  - Rating
  - Level of risk

Types of data: example

Name	Gender	Age	Marital status	No of children	Income	Smoking
John Smith	male	24	single	0	$25,000	never smoked
Mary Brown	female	35	married	3	$45,000	current smoker
Adam Jones	male	42	divorced	1	$40,000	former smoker
Jane Robertson	female	29	divorced	0	$42,000	never smoked
...	...	...	...		...	...

Uni-variate vs Multi-variate Data

VIGNAN'S
Foundation for Science, Technology & Research
(Deemed to be University)
-Estd. u/s 3 of UGC Act 1956

- Uni-variate data:
  - Data with single attribute/feature

Name	Income
John Smith	$25,000
Mary Brown	$45,000
Adam Jones	$40,000
Jane Robertson	$42,000
...	...

- Multi-variate data:
  - Data with single attribute/feature

Name	Gender	Age	Marital status	No of children	Income	Smoking
John Smith	male	24	single	0	$25,000	never smoked
Mary Brown	female	35	married	3	$45,000	current smoker
Adam Jones	male	42	divorced	1	$40,000	former smoker
Jane Robertson	female	29	divorced	0	$42,000	never smoked
...	...	...	...		...	...


SUMMARY

Thank You