WELLCOME

STATISTIC WITH PYTHON

EMBARKING ON A JOURNEY INTO DATA SCIENCE

YA MANON

COURSE OUTLINE

1	**Why Statistics?**	*Discuss the role of statistics in the context of **business intelligence** and **decision-making**, and **introduce the statistics workflow***
2	**Descriptive Statistics**	*Understand data using descriptive statistics, including **frequency distributions** and **measures of central tendency** , **variability**, **data visualization***
3	**Probability Distributions**	*Model data with probability distributions, and use the **normal distribution** to calculate probabilities and make value estimates*
4	**Central Limit Theorem**	*Introduce the **Central Limit Theorem**, which leverages the normal distribution to make inferences on populations with any distribution*
5	**Confidence Intervals**	*Make estimates with **confidence intervals**, which use sample statistics to define a range where an unknown population parameter likely lies*
6	**Hypothesis Tests**	*Draw **conclusions** with **hypothesis tests**, which let you evaluate assumptions about population parameters using sample statistics*
7	**Regression Analysis**	*Make **predictions** with **regression analysis**, and estimate the values of a dependent variable via its relationship with independent variables*

COURSE STRUCTURE

This is a **hands-on**, **project-based** course designed to help you apply statistical methods & techniques to real-world data analysis cases

Course resources include:

- **Downloadable PDF ebook** to serve as a helpful reference when you're offline or on the go (*or just need a refresher!*)
- **Quizzes** and **Project** to test and reinforce key concepts covered throughout the course, with detailed step-by-step solutions
- **Interactive demos** to keep you engaged, with downloadable Excel files, code that you can use to follow along from home

SETTING EXPECTATIONS

- The course simplifies essential statistics concepts for easy understanding
- No math or statistics background is required to take this course
- Focuses on the real-world application of statistical concepts
- Uses Microsoft Excel and Jupyter notebooks (Python) for hands-on demos and assignments
- Includes projects to test knowledge and apply it to different real-world scenarios

THE COURSE PROJECT

THE OBJECTIVES

- Understand the data with descriptive statistics
- Model the data with probability distributions
- Make estimates with confidence intervals
- Draw conclusions with hypothesis tests
- Make predictions with regression analysis

HELPFUL RESOURCES

Learn

Github: My Github account for code demos and excel file

E-learning

- *Sunrise E-Learning*
- *scribbr.com/category/statistics/*

YouTub

- My YouTube

Practice

Data Playground

- https://github.com/ManonYa09/Python_for_Data-Science-AI-/tree/main/Dataset

Online Datasets

- *kaggle.com/datasets*
- *data.world/datasets/open-data*
- *vincentarelbundock.github.io/Rdatasets/articles/data*
- *https://archive.ics.uci.edu/*

DATA SCIENCE

INTRO TO DATA SCIENCE

In this section we'll **introduce the field of data science**, discuss how it compares to other data fields, and walk through each phase of the data science workflow

TOPICS WE'LL COVER:

- What is Data Science
- Machine Learning
- Essential Skills
- Data Science Workflow

GOALS FOR THIS SECTION:

- Compare data science and machine learning with other common data analytics fields
- Introduce supervised and unsupervised learning, and examples of each technique
- Review the machine learning landscape and commonly used algorithms
- Discuss essential skills, and review each phase of the data science workflow

WHAT IS DATA SCIENCE?

- What is Data Science
- Essential Skills
- Machine Learning
- Data Science Workflow

Data science is about *using data to make smart decisions.*

Wait, isn't that **Business Interligence** ?

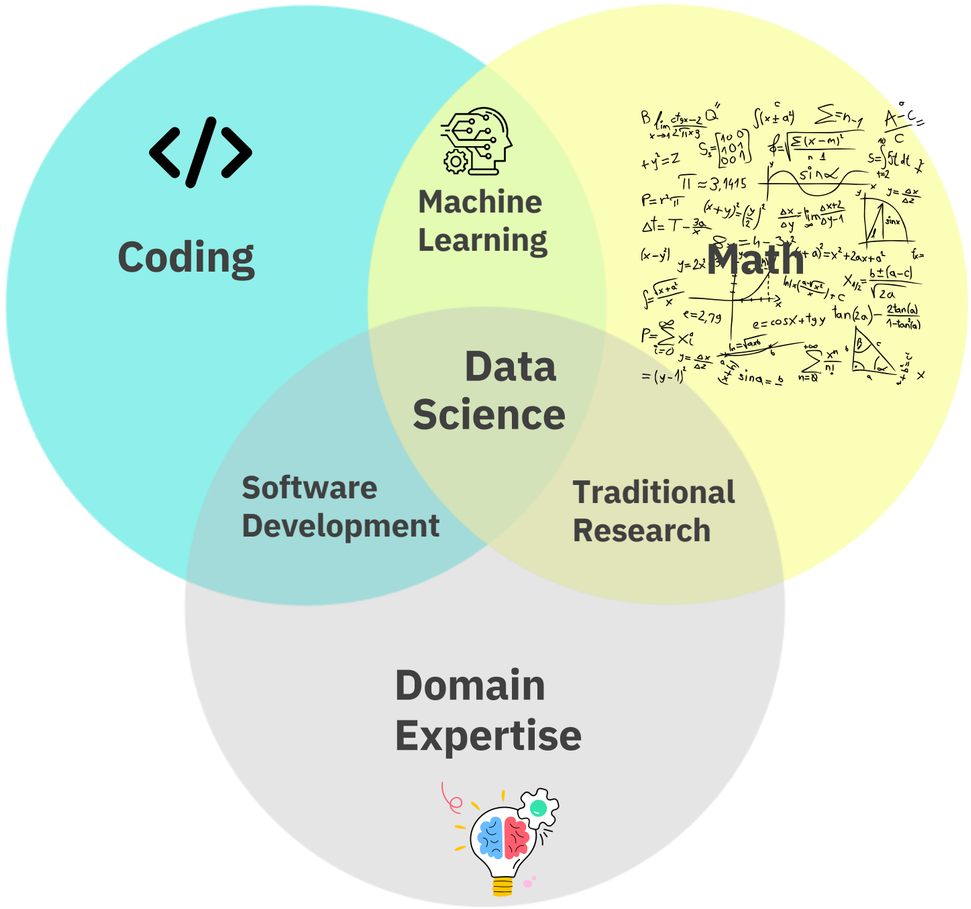
Yes! The differences lie in the **types of problems** you solve, and **tools and techniques** you use to solve them:

Business Interligence	Data Science
What *happened*?	**What's *going to* happen?**
• Descriptive Analytics • Data Analysis • Business Intelligence	• Predictive • Analytics • Data Mining

DATA SCIENCE SKILL SET

- What is Data Science
- Essential Skills
- Machine Learning
- Data Science Workflow

Data science requires a blend of **coding**, **math**, and **domain expertise**



The key is in applying these along with soft skills like:

- Communication
- Problem solving
- Curiosity & creativity
- Googling prowess

Data scientists & analysts approach problem solving in similar ways, but data scientists will often work with larger, more complex data sets and utilize advanced algorithms

WHAT IS MACHINE LEARNING?

What is Data Science

Essential Skills

Machine Learning

Data Science Workflow

Machine learning uses algorithms applied by data scientists to enable computers to learn and make decisions from data

Machine learning algorithms fall into two broad categories:

Supervised Learning

Using historical data to predict the future

What will house prices look like for the next 12 months?

How can I flag suspicious emails as spam?

Unsupervised Learning

Finding patterns and relationships in data

How can I segment my customers?

Which TV shows should I recommend to each user?

DATA SCIENCE WORKFLOW

- What is Data Science
- Essential Skills
- Machine Learning
- Data Science Workflow

The **data science workflow** consists of scoping the project, gathering, cleaning and exploring the data, applying models, and sharing insights with end users

This is not a linear process! *You'll likely go back to further gather, clean and explore your data*

WHY STATISTICS?

WHY STATISTICS FOR BUSINESS INTELLIGENCE?

In this section we'll discuss the **role of statistics** in the context of **business intelligence** and the decision-making process, review key terms, and introduce the statistics workflow

TOPICS WE'LL COVER:

Why Statistics

Populations

Statistics Workflow

GOALS FOR THIS SECTION:

- *Identify scenarios when statistics helps use data to make smart decisions, and when it's not needed*
- *Understand the concepts of populations & samples*
- *Review the statistics workflow and the concepts that will be covered throughout the course*

WHY STATISTICS?

Business intelligence is about using data to make **smart decisions**

Statistics is about *evaluating* those decisions under *uncertain* circumstances

When do you need statistics?

1) You don't have all the data you're interested in

- You can only analyze some of the data you need to make your decision
- There's **uncertainty** involved

2) The decision you're making is important

- You don't want to make the wrong one based on your limited data
- There's something specific to evaluate

What difference between BI and Data Science?

POPULATION & SAMPLES

Why Statistics

Populations

Statistics Workflow

A **population** contains all the data you're interested in to make your decision

- It's the data you wish you had, but are unlikely to get
- Any figure that summarizes a population is called a **parameter**

A **sample** contains some of the data from the population

- It's the data you have (which should ideally represent the population)
- Any figure that summarizes a sample is called a **statistic**

Statistics lets you make reasonable estimates about **parameters** using **statistics**

HEY THIS IS IMPORTANT!

Statistics can't create certainty out of uncertainty, it just helps you make controlled decisions under it!

THE STATISTICS WORKFLOW

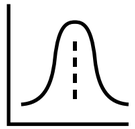
Why Statistics

Populations

Statistics Workflow

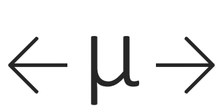
Descriptive Statistics

Understand what your sample data looks like



Probability Distribution

If the sample data fits a probability distribution, use it as a **model** for the entire population



Confidence Intervals

If the sample doesn't fit a distribution, use the central limit theorem to make **estimates** about population parameters

Hypothesis Testing

Continue to leverage the central limit theorem to draw **conclusions** about what a population looks like based on a sample

Regression Analysis

Use additional variables to increase the accuracy of your estimates and make **predictions** based on their relationships

HEY THIS IS IMPORTANT!

If you have all the population data, or simply need a bit of inspiration to make an "unimportant" decision, then descriptive statistics may be all you need!

DESCRIPTIVE STATISTICS

DESCRIPTIVE STATISTICS

In this section we'll cover understanding data with **descriptive statistics**, including frequency distributions, measures of central tendency, and measures of variability

TOPICS WE'LL COVER:

- **Statistics Basics**
- **Distributions**
- **Central Tendency**
- **Variability**

GOALS FOR THIS SECTION:

- *Identify the different types of variables in a dataset, along with their use cases*
- *Create frequency tables and plot the distributions of numerical variables using histograms*
- *Calculate the mean, median, mode, and standard deviation of a numerical variable*
- *Visualize the key descriptive statistics of a numerical variable using a box plot*

DESCRIPTIVE STATISTICS

Descriptive statistics consists of the collection, organization, summarization and presentation of data.

They reduce a large array of numbers into a handful of figures that describe it accurately

- Statistics Basics
- Distributions
- Central Tendency
- Variability

customer_id	age
15634602	42
15647311	41
15619304	42
15701354	39
15737888	43
15574012	44
15592531	50
15656148	29
15792365	44
15592389	27
15767821	31
15737173	24
15632264	34
15691483	25
15600882	35
15643966	45
15737452	58
15788218	24
15661507	45
15568982	24

n=10000