Introduction to Statistics

Purpose

The purpose of this lecture is to offer a brief overview of the field of Statistics which is paramount to performing Data Science. We will cover the most important introductory topics of Statistics starting with the type of statistics.

Purpose

The purpose of this lecture is to offer a brief overview of the field of Statistics which is paramount to performing Data Science. We will cover the most important introductory topics of Statistics starting with the type of statistics.

At the end of this lecture you will be able to:

1. Understand and define Statistics as well as differentiate between Descriptive and Inferential Statistics

What is Statistics

Statistics: Statistics is a branch of Mathematics that deals with collecting, organizing, and interpreting data

- Statistics is used in many fields from Social Science, Finance, Healthcare, and Sports.
- The goal of Statistics is to study a population and observe their characteristics. However, it is often difficult to obtain an entire population. Instead, we deal with a subset of the population known as a sample.

What is Statistics

Examples of areas you encounter in your day-to-day life which Statistics play a large role in:

1. Census Data
2. Sports Boxscores
3. Weather Forecasts
4. Political Campaigning
5. Stock Market

How is Statistics used in Data Science

Statistics is one of the main disciplines that make up the field of Data Science. Statistics provides tools for Data Scientists such as:

i. Determining how much of your result can be attributed to "Signal" and how much can be attributed to "Noise"
ii. Analyzing the efficacy of a data set and its collection methods before analysis is performed on the data

How is Statistics used in Data Science

Statistics provides tools for Data Scientists such as:

- Summarizing a data set in terms of descriptive statistics as well as plots and other metrics
- Testing a hypothesis one may have about the data (i.e. one variable influences the other, two groups are identical, etc.)
- Characterize data into one of the common distributions and then use this for prediction

Statistics in Python

Python has a wide range of useful functions to perform Statistical routines. These functions are found in the following two Modules:

- **scipy.stats**: The stats module in the SciPy Package offers many Statistical functions such as mean, z-score, correlation as well as other sub-modules for Continuous Distributions, Multivariate Distributions, and Discrete Distributions

Statistics in Python

Python has a wide range of useful functions to perform Statistical routines. These functions are found in the following two Modules:

- **numpy**: The numpy package itself offers many Statistical Functions such as Order Statistics, Averages and Variances, Correlating, and Histograms.

Statistics in Python

Example 1 (Importing Statistics Packages):

In [ ]:
```
import numpy as np
import pandas as pd
from scipy import stats
import seaborn as sns
import math
import random
import matplotlib.pyplot as plt
%matplotlib inline
```

In [ ]:
```
%whos
```

Statistics in Python

Example 2 (Testing Statistics Packages):

```
In [ ]: stats.describe([1,2,3,4,5]) # using scipy
```

```
In [ ]: np.mean([1,2,3,4,5]) # using numpy
```

Types of Statistics

There are two different types of statistics:

Descriptive Statistics: The purpose of Descriptive Statistics is to provide a summary of the data and its properties

Types of Statistics

Descriptive Statistics can take the form of simple metrics known as summary statistics (i.e. number of observations, minimum value, maximum value, mean, variance, etc.) or they can take the form of visualizations (i.e. histograms, scatterplots, pie charts, box plots, etc.)

Types of Statistics

Examples of Descriptive Statistics:

a. 75th percentile of height of men in the United States

b. Mean Field Goal Percentage of a professional basketball player in the NBA

c. Median salary of Data Scientist is across every major metropolitan area in the United States

Types of Statistics

Inferential Statistics: The purpose of Inferential Statistics is to make inferences about a population using a subset of the population known as a sample

Inferential Statistics begin with a hypothesis about the population and then the sample is used to prove or disprove the hypothesis, effectively inferring something about the population

Types of Statistics

Examples of Inferential Statistics:

- Survey is sent out to 1000 residents in Chicago asking them about their political views. The survey designers are interested in knowing if Chicago residents in different areas have more or less conservative views.

Types of Statistics

Examples of Inferential Statistics:

- A clinician develops a new drug to help patients relieve anxiety. The clinician collects a sample of individuals and gives half of them the new drug and the other half are given a placebo drug

Helpful Points

- It is important to understand that the purpose of Descriptive Statistics is to simply collect and record metrics and nothing more. It does NOT involve any generalization beyond the summary statistics
- In an experiment it is common to use both Descriptive AND Inferential Statistics