**Descriptive Statistics- A Primer**

**Krishna Gopal Goswami, PGDM – Business Analytics**

Contents

[1. Introduction to Descriptive Statistics 3](#_Toc505376814)

[1.1 Types of Statistics 3](#_Toc505376815)

[1.2 Types of Data 4](#_Toc505376816)

[1.3 Types of Analysis 4](#_Toc505376817)

[1.4 Descriptive Statistics Classifications 5](#_Toc505376818)

[1.4.1 Measures of Central Tendency 5](#_Toc505376819)

[1.4.2 Measures of Dispersion/Variability 7](#_Toc505376820)

[1.4.3 Measures of Shape 9](#_Toc505376821)

[1.4.4 Graphical Methods for displaying data 12](#_Toc505376822)

[2. Limitation/Challenges 18](#_Toc505376823)

[3. Case Study 18](#_Toc505376824)

[4. Annexure 22](#_Toc505376825)

[5.1 Packages Used 23](#_Toc505376826)

[5.2 Link to Github 23](#_Toc505376827)

[5.3 References 23](#_Toc505376828)

# Introduction to Descriptive Statistics

**What is Statistics?**

**Statistics** is a branch of mathematics dealing with the collection, analysis, interpretation, presentation and organization of data.

# Types of Statistics

1. **Descriptive Statistics**: Descriptive Statistics is the branch of statistics, which is concerned with describing the population under study.
2. **Inferential Statistics:** Inferential Statistics is a type of statistics, which focuses on drawing conclusions about the population, on the basis of sample analysis and observation.

**Difference between Descriptive and Inferential Statistics**

|  |  |  |
| --- | --- | --- |
| **Basis for comparison** | **Descriptive Statistics** | **Inferential Statistics** |
| What does it do? | Organize, analyse and present data in a meaningful way. | Compares, test and predicts data. |
| Form of Final Result | Charts, Graphs and Tables | Probability |
| Usage | To describe a situation. | To explain the chances of occurrence of an event. |
| Function | It explains the data, which is already known, to summarize sample. | It attempts to reach the conclusion to learn about the population, which extends beyond the data available. |

# Types of Data

There are 2 broad specifications of data:

1. **Categorical (Qualitative) Data**: It describes the quality of the variable/attribute. is useful for classifying data into groups or classes. This type of data cannot be measured. e.g. Gender (Male/Female)
2. **Quantitative Data**: Data that can be measured is called Quantitative data. The possibility of values for this attribute is unlimited. e.g. Height, Weight, Temperature

There are 2 types of Categorical data:

1. **Nominal Data:** No inherent order in categories. e.g. Region, Gender etc.
2. **Ordinal Data:** Categories have inherent ordering. e.g. Job Grade etc.

There are 2 types of Quantitative data:

1. **Interval Data:** Difference between any 2 values makes sense, but the division of values does not fetch any meaningful result. E.g Salary Ranges, Dates, Height etc.
2. **Ratio Data:** Both Difference and Division fetch meaningful result. E.g. BMI = Height/Weight

# Types of Analysis

1. **Univariate**: One variable is analysed at a time. The objective is to describe the variable. Example- How many students are graduating with “Analytics“ degree?
2. **Bivariate**- Two variables are analysed together for any possible association or empirical relationship. Example- What is the correlation between “Gender” and graduation with “Analytics” degree?
3. **Multivariate**- More than two variables are analysed together for any possible association or interactions. Example – What is a correlation between “Gender”, “Country of Residence” and graduation with “Analytics” degree?

# Descriptive Statistics Classifications

Descriptive Statistics can be classified into 3 categories:

* + **Measures of Central Tendency**
  + **Measures of Dispersion/Variability**
  + **Measures of shape**

# Measures of Central Tendency

It seeks to characterize the distribution's central or typical value.

There are 3 ways to calculate measures of central tendency;

* Mean
* Median
* Mode

**Mean**

The mean is the most popular and well-known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.

The mean is equal to the sum of all the values in the data set divided by the number of values in the dataset. So, if we have n values in a data set and they have values x1, x2, ..., xn, the **sample mean**, usually denoted by https://statistics.laerd.com/statistical-guides/img/measures-of-central-tendency-2.png (pronounced x bar), is:

https://statistics.laerd.com/statistical-guides/img/measures-of-central-tendency-1.png

This is also represented as https://statistics.laerd.com/statistical-guides/img/measures-of-central-tendency-4.png

To represent Population Mean, we use Greek Letter µ (mu):

https://statistics.laerd.com/statistical-guides/img/measures-of-central-tendency-5.png

When not to use Mean:

1. It is particularly susceptible to the influence of outliers
2. We usually prefer the median over the mean (or mode) is when our data is skewed.

**Median**

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data.

Disadvantages of using Median:

It does not take all the values into account. It is based on only the middle value by **virtue of its position in the centre.**

**Mode**

The mode is the most frequent score in our data set. Normally, the mode is used for categorical data where we wish to know which the most common category.

Disadvantages of using Mode:

1. It will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the dataset.
2. This is particularly problematic when we have continuous data because we are more likely not to have anyone value that is more frequent than the other.

**Summary of when to use the mean, median and mode:**

Please use the following summary table to know what the best measure of central tendency is with respect to the different types of variable.

|  |  |
| --- | --- |
| **Type of Variable** | **Best measure of central tendency** |
| Nominal | Mode |
| Ordinal | Median |
| Interval/Ratio (not skewed) | Mean |
| Interval/Ratio (skewed) | Median |

# Measures of Dispersion/Variability

These statistics describe how the data varies or is dispersed (spread out).

Main measures of Dispersion are:

* Range
* Standard Deviation
* Variance
* Interquartile Range

**Range:**

Range is the difference between the highest and the lowest value in the data values.

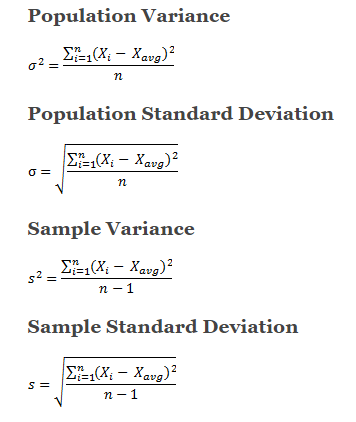
It is mostly used to find out what band the values of the data fall under. Ex: salary range, age group.

**Variance and Standard Deviation:**

Variance is the sum of all the squared deviations from the mean by N.

Standard Deviation is the square root of the variance.

Standard Deviation is the most popular measure of variability



**Interquartile Range**

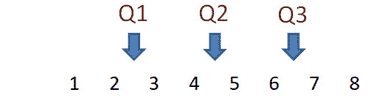
It is a measure of variability, based on dividing a data set into quartiles.

Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.

Here's how to find the IQR:

**Step 1:** Put the data in order from least to greatest. e.g. Consider the following set of numbers which are already ordered from least to greatest

1, 2, 3, 4, 5, 6, 7, 8



**Step 2:** Find the median. If the number of data points is odd, the median is the middle data point. If the number of data points is even, the median is the average of the middle two data points.

Here Median will be: Q2 = (4 + 5)/2 = 4.5

**Step 3:** Find the first quartile (Q​1​​). The first quartile is the median of the data points to the left of the median in the ordered list.

Here Q1 will be (1+2)/2 = 2.5

**Step 4:** Find the third quartile (Q​3​​).The third quartile is the median of the data points to the right of the median in the ordered list.

Here Q3 will be (6+7)/2 = 6.5

**Step 5:** Calculate IQR by subtracting Q3 – Q1.

IQR is 6.5 – 2.5 = 4

# Measures of Shape

Measures of shape describe the distribution (or pattern) of the data within a dataset**.**

There are 2 measures of shapes:

1. Skewness
2. Kurtosis

**Skewness:**



Skewness is when a distribution is asymmetrical or lacks symmetry, the data is said to be skewed either rightwards or leftwards. The skewed portion is the long thin part of the curve.

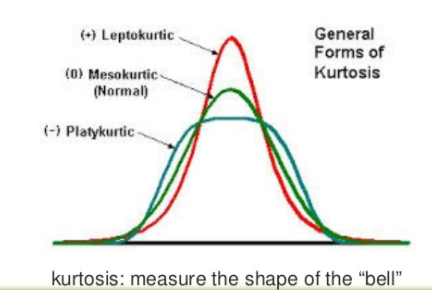
If the bell curve is to the right and the skew tail is to the left, then it is NEGATIVELY SKEWED. If the skewed portion is to the right and the curve is to the left, it is POSITIVELY SKEWED.

**Kurtosis:**

**Kurtosis tells you how tall and sharp the central peak is, relative to a standard bell curve.**

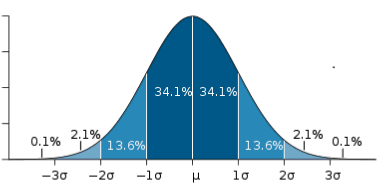
There are 3 forms of Kurtosis:

1. Leptokurtic: A distribution with kurtosis>3. Its central peak is higher and sharper.
2. Mesokurtic: A normal distribution has kurtosis =3.
3. Platykurtic: A distribution with kurtosis<3. Its central peak is lower and broader.



**Normal Distribution**

A normal distribution is a true symmetric distribution of observed values

If represented as a 'normal curve' (or bell curve) the graph would take the following shape (where µ = mean, and σ = standard deviation):  
  
  
Key features of the **normal distribution**:

* symmetrical shape
* [Mode](http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+statistical+language+glossary#Mode), [Median](http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+statistical+language+glossary#Median) and [Mean](http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+statistical+language+glossary#Mean) are the same and are together in the centre of the curve
* There can only be one mode (i.e. there is only one value which is most frequently observed)
* Most of the data are clustered around the centre, while the more extreme values on either side of the centre become less rare as the distance from the centre increases (i.e. About 68% of values lie within one standard deviation (*σ*) away from the mean; about 95% of the values lie within two [standard deviations](http://www.abs.gov.au/websitedbs/a3121120.nsf/home/statistical+language+-+statistical+language+glossary#Standard%20deviation), and about 99.7% are within three standard deviations. This is known as the *empirical rule* or the *3-sigma rule*.)

**Outliers**:

Outliers are values that are far from the central tendency. Outliers might

be caused by errors in collecting or processing the data, or they might be

correct but unusual measurements.

# Graphical Methods for displaying data

1. Frequency Distribution
2. Histogram
3. Scatter Plots
4. Pareto Charts
5. Boxplots

**Frequency Distribution**

 A **frequency distribution** is a table or graph that displays the frequency of various outcomes in a sample.

Each entry in the table contains the [frequency](https://en.wikipedia.org/wiki/Frequency_(statistics)) or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the [distribution](https://en.wikipedia.org/wiki/Statistical_distribution) of values in the sample.

Here is an example of a univariate (i.e. single [variable](https://en.wikipedia.org/wiki/Variable_(mathematics))) frequency table. The frequency of each response to a survey question is depicted.

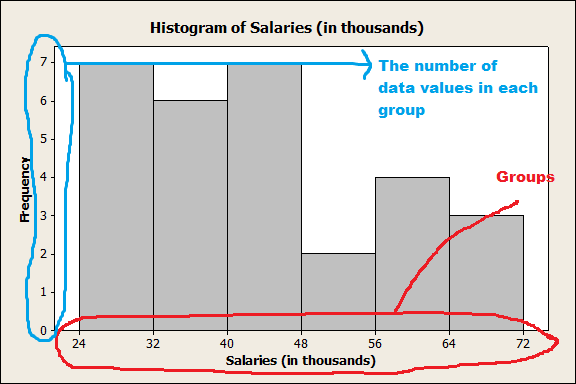
|  |  |  |
| --- | --- | --- |
| [**Rank**](https://en.wikipedia.org/wiki/Ranking) | **Degree of agreement** | **Number** |
| 1 | Strongly agree | 20 |
| 2 | Agree somewhat | 30 |
| 3 | Not sure | 20 |
| 4 | Disagree somewhat | 15 |
| 5 | Strongly disagree | 15 |

**Histogram:**

A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of [continuous](https://statistics.laerd.com/statistical-guides/types-of-variable.php) data.

This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc.

**The general idea behind a histogram is to divide the data set into groups of equal length** which allows us to see the patterns in the data instead of the detailed information we would get from what is basically a list of numbers.



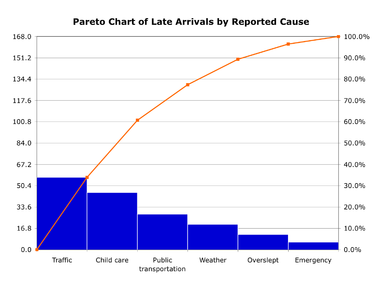
**Pareto Chart**

It is a type of chart that contains both bars and a line graph, where individual values are represented in descending order by bars, and the cumulative total is represented by the line.

It helps to determine which values are least important and which values are more important.

When to use Pareto Chart:

* When analysing data about the frequency of problems or causes in a process.
* When there are many problems or causes and you want to focus on the most significant.
* When analyzing broad causes by looking at their specific components.



**Scatter Plot**

Scatter plot looks at the relationship between two or more variables. It is a great way to identify outliers.

When to use scatter plot

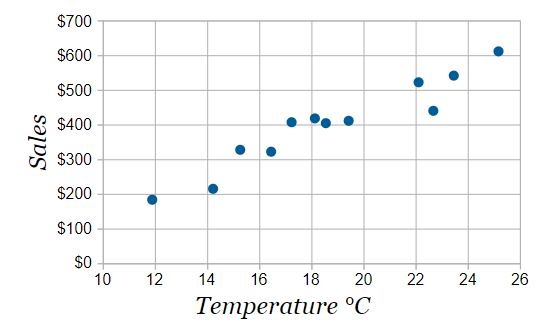
* When you have paired numerical data.
* When trying to determine whether the two variables are related,

Example

The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days:

|  |  |
| --- | --- |
| ***Ice Cream Sales vs Temperature*** | |
| **Temperature °C** | **Ice Cream Sales** |
| 14.2° | $215 |
| 16.4° | $325 |
| 11.9° | $185 |
| 15.2° | $332 |
| 18.5° | $406 |
| 22.1° | $522 |
| 19.4° | $412 |
| 25.1° | $614 |
| 23.4° | $544 |
| 18.1° | $421 |
| 22.6° | $445 |
| 17.2° | $408 |

* And here is the same data as a Scatter Plot:



* It is now easy to see that **warmer weather leads to more sales**, but the relationship is not perfect.

**Box and Whisker Plot**

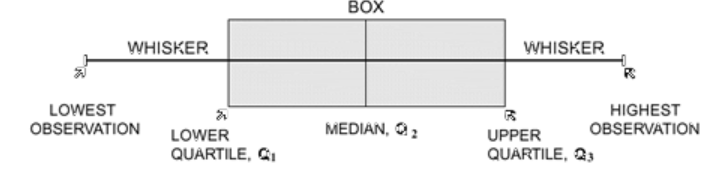
**Box and Whisker Plot (**also called **box plot)** is a method for graphically depicting groups of numerical data through their quartiles.

When to use Box Plot:

1. It is ideal for comparing distributions because the centre, spread and overall range are immediately apparent.
2. It is a way of summarizing a set of data measured on an interval scale. It is often used in explanatory data analysis.
3. This type of graph is used to show the shape of the distribution, its central value, and its variability.

In a box and whisker plot:

* The ends of the box are the upper and lower quartiles, so the box spans the interquartile range
* The median is marked by a vertical line inside the box
* The whiskers are the two lines outside the box that extend to the highest and lowest observations.



# Limitation/Challenges

Descriptive statistics are limited in so much that they only allow you to make summations about the people or objects that you have actually measured.

You cannot use the data you have collected to generalize to other people or objects (i.e., using data from a sample to infer the properties/parameters of a population).

 For example, if you tested a drug to beat cancer and it worked in your patients, you cannot claim that it would work in other cancer patients only relying on descriptive statistics (but inferential statistics would give you this opportunity).

# Case Study

**Problem Statement:**

We will perform exploratory, diagnostic and descriptive statistics in the iris dataset as part of Exploratory Data Analysis using Python.

**Exploratory Data Analysis:**

1. **Load the required libraries**

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

%matplotlib inline

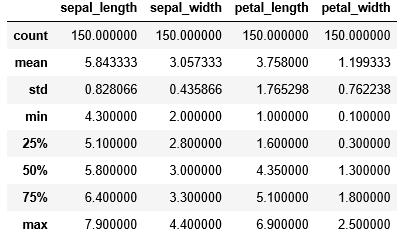
plt.show()

1. **Perform basis statistic on iris dataset**

**Command**:

iris.describe()

**Result**:



1. **Analyse distribution of the dataset using histogram and Kernel Density Estimator**

Kernel density estimation is a fundamental data smoothing problem where inferences about the [population](https://en.wikipedia.org/wiki/Statistical_population) are made, based on a finite data [sample](https://en.wikipedia.org/wiki/Statistical_sample).

**Command:**

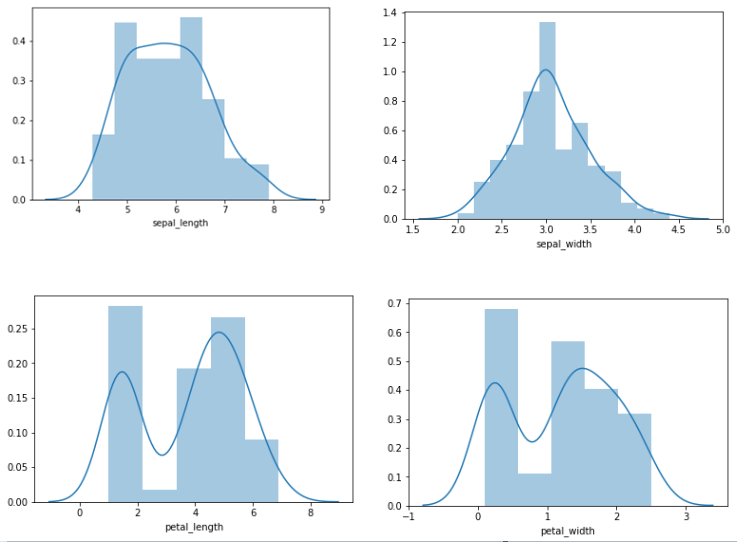
sns.distplot(iris.sepal\_length)

sns.distplot(iris.sepal\_width)

sns.distplot(iris.petal\_length)

sns.distplot(iris.petal\_width)

**Result:**



**From the above result, we can see that:**

Petal Length and Petal Width do not follow Normal Distribution.

Sepal Length and Sepal Width follow Normal Distribution.

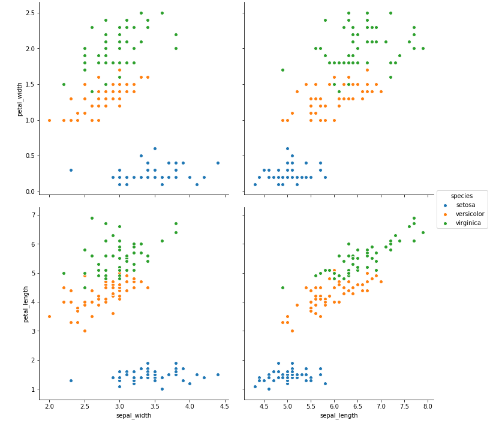
1. Run pair plot of the seaborn library to plot pairwise relationships in a dataset.

**Command:**

sns.pairplot(iris, hue = "species", size =5, x\_vars = ["sepal\_width", "sepal\_length"], y\_vars = ["petal\_width", "petal\_length"])

plt.show();

**Result**:



**From the above result, we can see that** :

Iris-setosa is very distinctly different from others.

Iris-versicolor and iris-virginica can be differentiated based on petal\_length and petal\_width.

1. **Run box plot, to see the distribution pattern and check outlier.**

**Command:**

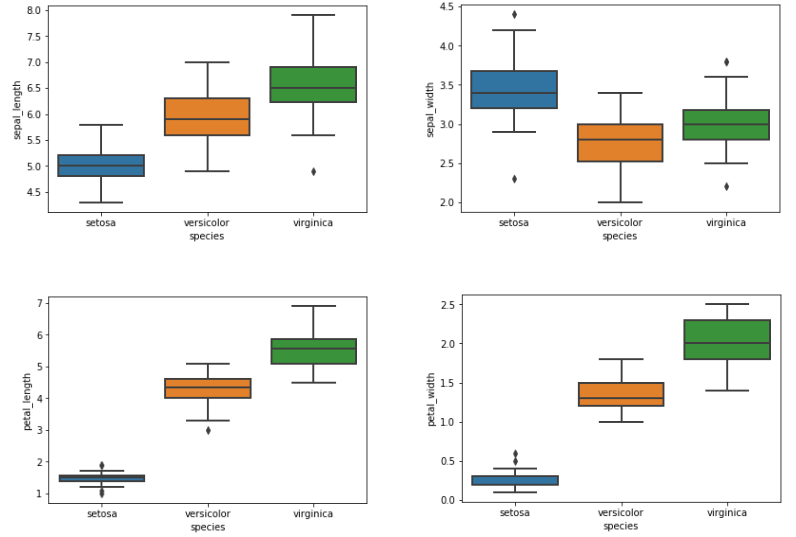
sns.boxplot(x='species', y='sepal\_length', data = iris, linewidth = 2)

sns.boxplot(x='species', y='sepal\_width', data = iris, linewidth = 2)

sns.boxplot(x='species', y='petal\_length', data = iris, linewidth = 2)

sns.boxplot(x='species', y='petal\_width', data = iris, linewidth = 2)

**Result:**



From the above result, we can see that,

There are outliers in some cases

Medians vary for each of the species.

# Annexure

# Packages Used

The following packages in Python code are used for the descriptive and exploratory analysis.

**numpy**: Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**seaborn**: Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics.

**matplotlib:** Matplotlib is a plotting library for the Python programming language. pyplot is a matplotlib module which provides a MATLAB-like interface

# Link to Github

The following Github link contains Python code used in Case Study in section 4.

# References

1. https://keydifferences.com/difference-between-descriptive-and-inferential-statistics.html
2. <https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php>
3. <https://datafai.com/>
4. <https://www.slideshare.net/doujou.DC/descriptive-statistics-and-data-visualization>
5. https://www.kaggle.com/danalexandru/simple-analysis-of-iris-dataset