­CCT College Dublin

|  |  |
| --- | --- |
| Module Titles: | Data Preparation & Visualisation, Machine Learning for Data Analytics, Programming for Data Analytics, Statistics for Data Analytics |
| Assessment Title: | Integrated CA, Ireland Tourism |
| Lecturer Names: | David McQuaid, Dr. Muhammad Iqbal, Sam Weiss, Taufique Ahmed |
| Student Full Name: | Ricardo Alves de Souza |
| Student Number: | 2018499 |
| Assessment Due Date: | 05/04/2024 |
| Date of Submission: | 05/04/2024 |

Declaration

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Title: Integrated CA, Ireland Tourism

Author: Ricardo Alves de Souza

E-mail: [2018499@student.cct.ie](mailto:2018499@student.cct.ie?subject=Integrated%20CA1)

GitHub Address: <https://github.com/ricardoasouz/2024---MSc-in-Data-Analytics---Feb---FT/tree/main/MSC_DA_CA1>

Table of Contents

[Abstract 5](#_Toc162372712)

[Introduction 5](#_Toc162372713)

[Difference between Discrete Data and Continuous Data. 6](#_Toc162372714)

[Definition of discrete data. 6](#_Toc162372715)

[Definition of continuous data. 6](#_Toc162372716)

[Definition of Descriptive Statistics 6](#_Toc162372717)

[Common descriptive statistics in general include: 6](#_Toc162372718)

[Some common formulas and calculations used in Descriptive Statistics 8](#_Toc162372719)

[Minimum 8](#_Toc162372720)

[Maximum 8](#_Toc162372721)

[Range 8](#_Toc162372722)

[Sum 8](#_Toc162372723)

[Size / Count 9](#_Toc162372724)

[Mean 9](#_Toc162372725)

[Median 9](#_Toc162372726)

[Mode 10](#_Toc162372727)

[Standard Deviation 10](#_Toc162372728)

[Variance 11](#_Toc162372729)

[Midrange 11](#_Toc162372730)

[Quartiles 11](#_Toc162372731)

[Interquartile Range 12](#_Toc162372732)

[Outliers 12](#_Toc162372733)

[Sum of Squares 12](#_Toc162372734)

[Mean Absolute Deviation 13](#_Toc162372735)

[Root Mean Square 13](#_Toc162372736)

[Standard Error of the Mean 13](#_Toc162372737)

[Skewness 14](#_Toc162372738)

[Kurtosis 14](#_Toc162372739)

[Coefficient of Variation 15](#_Toc162372740)

[Relative Standard Déviation 16](#_Toc162372741)

[Frequency 16](#_Toc162372742)

[Reference list: 17](#_Toc162372743)

­

[Figure 1 - Table of Some Descriptive Statistics 7](#_Toc162356614)

­­

# Abstract

Times New Roman 12-point italics.

# Introduction

# Difference between Discrete Data and Continuous Data.

## Definition of discrete data.

## Discrete data refers to a form of quantitative information characterized by countable figures and non-fractional values. Typically, discrete data is presented in the form of whole numbers that convey precise quantities. A common way to conceptualize discrete data is to preface it with "the number of," for instance, the number of patrons in a shop. This kind of data generally encapsulates distinct occurrences that are already in the past. In analysing discrete data, you can examine precise numbers, such as the quantity of products sold on a particular date or the duration of time an employee has worked in a given week.

## Definition of continuous data.

Continuous data is a quantitative data category that captures measurements that can be highly precise, extending to numerous decimal places as needed. It represents values that can be measured on a scale and can fall between any two amounts within a range. This data type is prevalent in sectors that demand exactness, such as healthcare, production, and research and development. Continuous data is dynamic, presenting the opportunity for organizations to scrutinize their processes and forecast upcoming patterns. An instance of its application could be a company monitoring the duration required by a team to fulfil assignments, providing insights into productivity and efficiency.

Top of Form

Bottom of Form

# 

# Definition of Descriptive Statistics

Descriptive statistics provide a numerical summary of the characteristics of a collected dataset, a population, or a subset thereof. These calculations are designed to convey the central tendency, dispersion, and shape of the dataset’s distribution.

## Common descriptive statistics in general include:

* Count / Size
* Minimum
* Maximum
* Sum
* Mean
* Median
* Average
* Midrange
* Standard deviation
* Quartiles

But this is not all, descriptive statistics can have more different types of analyses, like and maybe more than have in the following table:

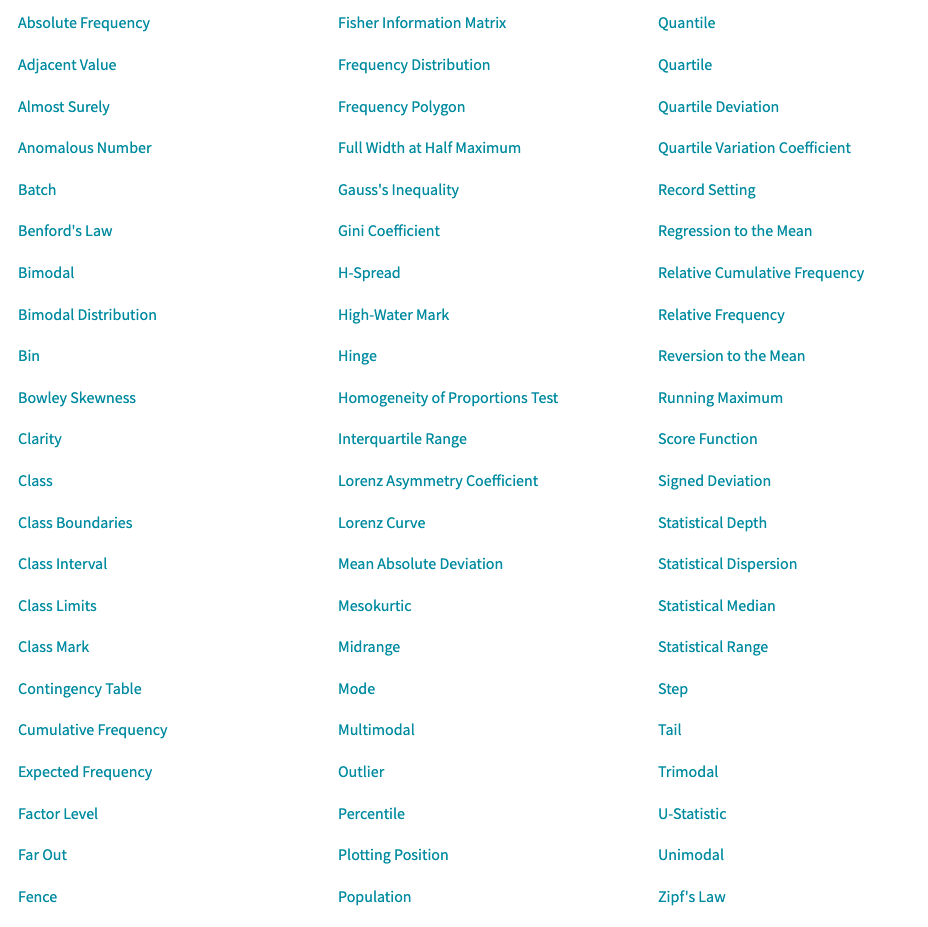


Figure 1 - Table of Some Descriptive Statistics

Descriptive statistics from Wolfram MathWorld. Available at: https://mathworld.wolfram.com/topics/DescriptiveStatistics.html (Accessed: 26 March 2024).

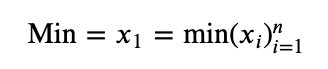
# Some common formulas and calculations used in Descriptive Statistics

## Minimum

Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn

The minimum in a data set is the least value​, when the data is arranged in ascending order from the smallest to the largest value

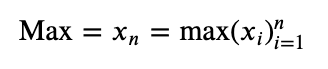


## Maximum

The maximum in a dataset represents the greatest value located at the far right when the data is ordered from the lowest to the highest value.

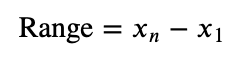
Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn



Range

The difference between the minimum and maximum values in a dataset is known as the range. It is calculated by subtracting the minimum value, from the maximum value. The range provides a measure of the spread or dispersion of the data points within the set.

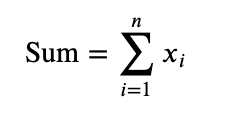


## 

## Sum

The total of all data values in a dataset is known as the sum of the aggregate. It is calculated by adding together all the individual values in the dataset. This total is often symbolized by the Greek letter Sigma (Σ) followed by the expression for the data points, indicating the summation of the series of values.

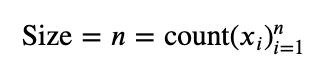
x1 + x2 + x3 + ... + xn



## Size / Count

The size or count of a dataset refers to the number of individual data points it contains. This is a measure of the dataset's magnitude in terms of its elements and is often denoted as *n* in statistical notation.

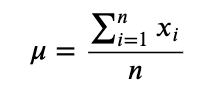
Top of Form

Bottom of Form

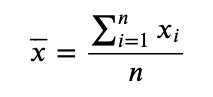
## Mean

The mean, also commonly referred to as the average, is a measure of the central tendency of a dataset. It is calculated by adding all the data values together to find the sum and then dividing this total by the number of data points in the set, which is the size or count.

For a Population:



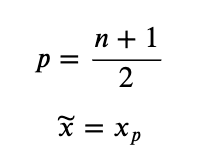
For a Sample:



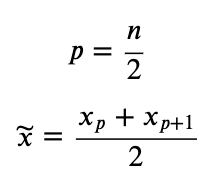
## Median

When arranging a dataset from smallest to largest, the median is the middle value that divides the dataset into two halves. For datasets with an odd number of entries, the median is the central value. However, if the dataset has an even number of entries, the median is found by calculating the average of the two middle values.

If n is an odd number:



If n is even the median is the average of the values at positions p and p + 1 where



## Mode

The mode in a dataset represents the value(s) that occur most frequently. A dataset can have one mode and more than one mode if multiple values occur with the same highest frequency, or no mode if all values are unique.

Top of Form

Bottom of Form

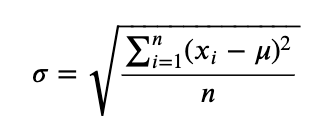
Top of Form

Bottom of Form

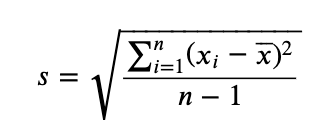
## Standard Deviation

Standard deviation quantifies how spread out the data points in a set are from their average value. This measure is calculated by taking the square root of the average of the squared deviations of each data point from the mean. A larger standard deviation indicates that the data points are more widely dispersed from the mean.Top of Form

For a Population



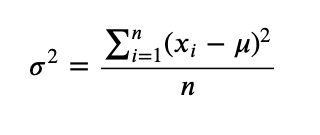
For a Sample



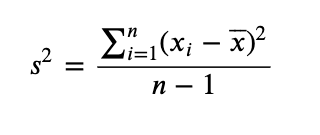
## Variance

Variance measures the spread of data points around the mean within a dataset. It's computed as the average of the squared deviations from the mean. A greater variance indicates that the data points are more dispersed from the mean, highlighting the variability within the dataset.

For a Population



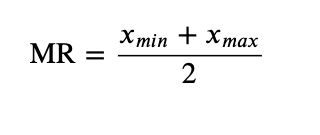
For a Sample



## Midrange

## 

The midrange in statistics is found by averaging the smallest and largest numbers in a dataset. It provides a quick sense of the centre or middle value of the data, especially useful for understanding the data's range.



## Quartiles

Quartiles divide a dataset into four equal parts.

The median, or second quartile (Q2), splits the data into upper and lower halves.

The first quartile (Q1) is the median of the data points below Q2

The third quartile (Q3) is the median of the data points above Q2.

These quartiles help in understanding the distribution of data by highlighting the *spread* and *central tendency*.

\**This is one of several methods for calculating quartiles.*

## Interquartile Range

The interquartile range (IQR) is defined as the distance between the first quartile (Q1) and the third quartile (Q3) in a dataset. It represents the range within which the middle 50% of the data points lie, effectively measuring the spread of the central portion of the dataset and minimizing the impact of *outliers*.

𝐼𝑄𝑅=𝑄3−𝑄1

## Outliers

Potential outliers are those values in a dataset that fall either below the Lower Fence or above the Upper Fence. These fences are determined by specific calculations that take into account the interquartile range (IQR), helping to identify data points that significantly differ from the rest of the dataset.

Upper Fence = 𝑄3+1.5×𝐼𝑄𝑅

Lower Fence = 𝑄1−1.5×𝐼𝑄𝑅

## Sum of Squares

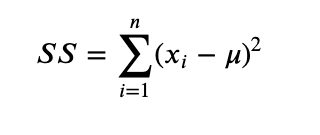
The sum of squares refers to the aggregated total of each data point's deviation from the mean, squared.

This calculation is a fundamental part of various statistical analyses, serving to quantify the variance within a dataset.

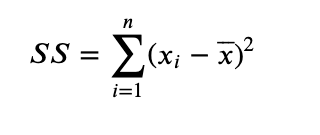
Top of Form

Bottom of Form

For a Population



For a Sample

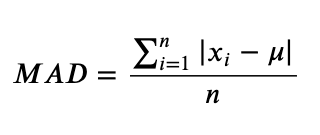


## 

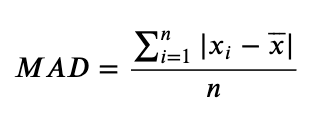
## Mean Absolute Deviation

Mean absolute deviation measures the average distance between each data point and the mean of the dataset. This is calculated by taking the absolute values of the differences between each data point and the dataset's mean, and then dividing by the number of data points. It provides insight into the variability of the dataset.

For a Population

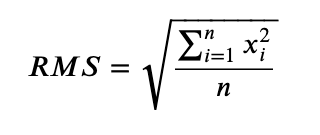


For a Sample



## Root Mean Square

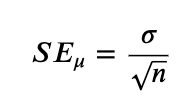
The root mean square (RMS) is a statistical measure that calculates the magnitude of a set of numbers. It is found by taking the square root of the average of the squares of the values in the set. This metric is especially useful in contexts where both positive and negative values in the dataset are treated equally, and it tends to give a higher value than the average due to the squaring of the values.



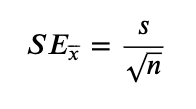
## Standard Error of the Mean

The standard error of the mean (SEM) is derived by dividing the standard deviation of the dataset by the square root of the number of observations (*n*). This metric indicates how much the sample mean is expected to vary from the true population mean.

For a Population



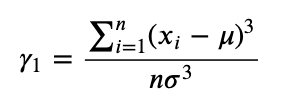
For a Sample



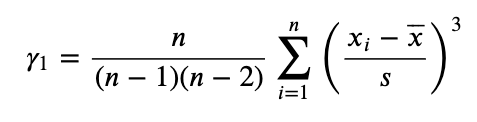
## Skewness

Skewness measures the asymmetry of a data distribution compared to the normal distribution. If the distribution has a longer tail on the left side, it is considered left-skewed or negatively skewed. Conversely, if it has a longer tail on the right side, it's right-skewed or positively skewed. This characteristic helps in understanding the direction and extent of distribution deviation from the symmetrical bell curve.

For a Population



For a Sample



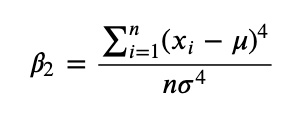
## Kurtosis

Kurtosis measures the "tailedness" of a distribution, indicating how outlier-prone a dataset is. High kurtosis suggests more extreme outliers than a normal distribution, while low kurtosis indicates fewer extreme outliers. This helps assess the extremity and concentration of tail data compared to a normal bell curve.

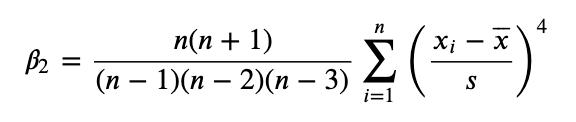
Top of Form

Bottom of Form

For a Population



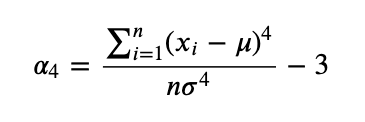
For a Sample



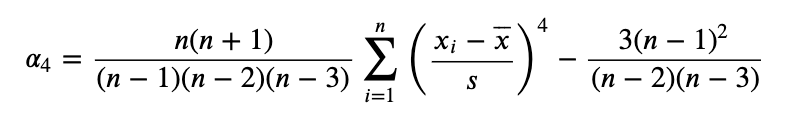
Kurtosis Excess

Excess kurtosis gauges the peak height of a distribution's tails, focusing on the concentration of outliers rather than their extremity. A distribution with high excess kurtosis indicates a significant presence of outlier data, pointing to more frequent extreme deviations from the mean compared to a normal distribution.

For a Population



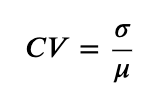
For a Sample #(This is just Kurtosis in MS Excel and Google Sheets)



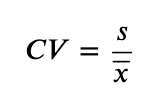
## Coefficient of Variation

The coefficient of variation (CV) measures the relative spread of data points around the mean, expressed as a ratio of the standard deviation to the mean. It's a useful statistic for comparing the degree of variability from one data series to another, even if the means are drastically different. The CV is calculated by dividing the standard deviation by the mean. This measure is particularly helpful in assessing the risk or variability in different contexts, such as finance and scientific research, where understanding relative dispersion is crucial.

For a Population



For a Sample



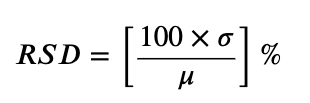
## Relative Standard Deviation

The relative standard deviation (RSD) quantifies the variation in a data set relative to its mean, presented as a percentage. It's computed by multiplying the standard deviation by 100 and then dividing by the mean. This statistic is valuable for comparing the variability of datasets with different units or means, providing a normalized measure of dispersion.

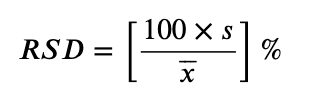
Top of Form

Bottom of Form

For a Population



For a Sample



## Frequency

Frequency measures how often each value appears in a dataset, essential for determining the mode, the value that occurs most frequently. This statistical concept helps in understanding the distribution and concentration of data points.

Top of Form

Bottom of Form

# Reference list

*Descriptive statistics* *from Wolfram MathWorld*. Available at: https://mathworld.wolfram.com/topics/DescriptiveStatistics.html (Accessed: 26 March 2024).