­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:

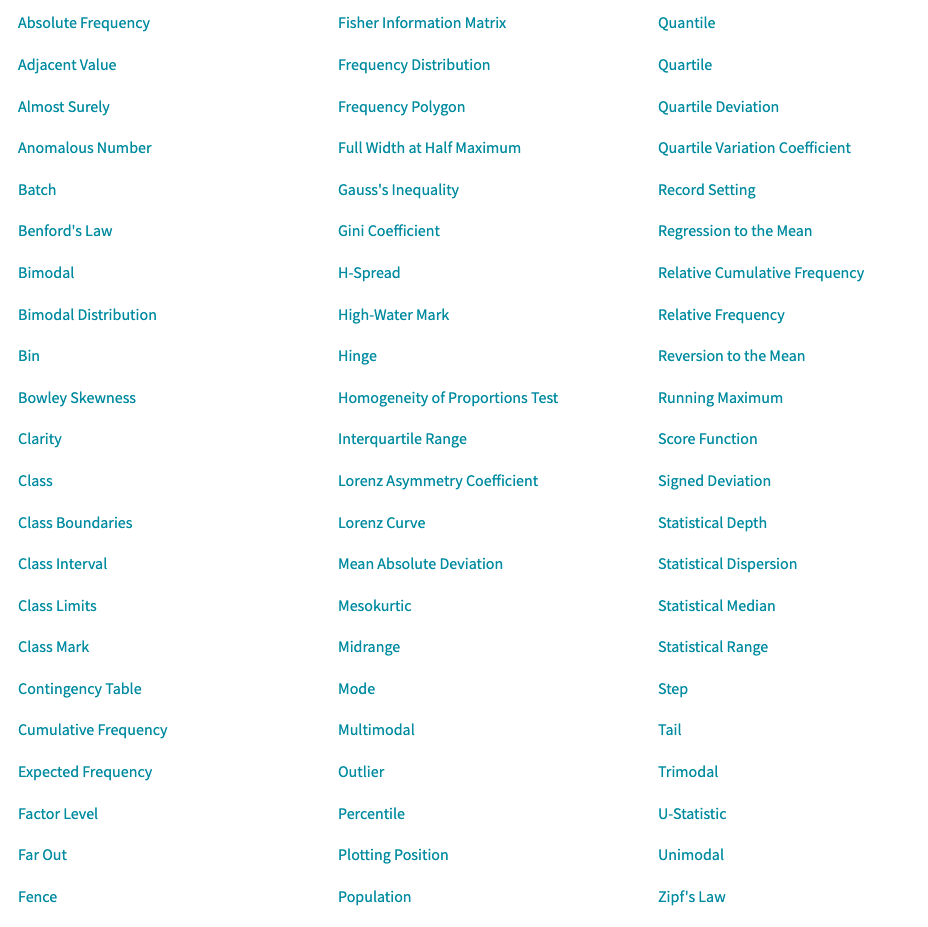


Figure 1 - Table of Some Descriptive Statistics

Descriptive statistics (no date) from Wolfram MathWorld. Available at: https://mathworld.wolfram.com/topics/DescriptiveStatistics.html (Accessed: 26 March 2024).

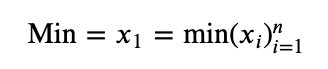
# Some common formulas and calculations used in Descriptive Statistics

## 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

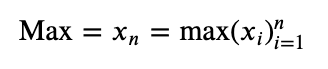


## Maximum

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.

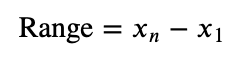
Ordering a data set:

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.

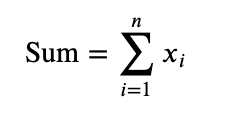


## 

## 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.

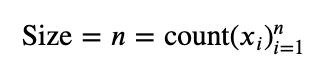
x1 + x2 + x3 + ... + xn



## 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.

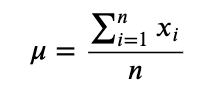
Top of Form

Bottom of Form

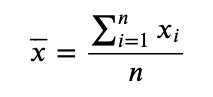
## Mean

The mean, also commonly referred to as the average, is a measure of 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. The formula for the mean is typically expressed as �ˉ=∑���*x*ˉ=*n*∑*xi*​​, where ∑��∑*xi*​ represents the sum of all data values and �*n* is the count of data points.

For a Population:



For a Sample:

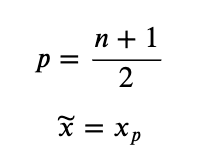


## Median

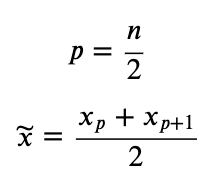
Ordering a data set x1 ≤ x2 ≤ x3 ≤ ... ≤ xn from lowest to highest value, the median is the numeric value separating the upper half of the ordered sample data from the lower half.

If n is odd the median is the center value. If n is even the median is the average of the 2 center values.

If n is odd the median is the value at position p where.



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



## Mode

The mode is the value or values that occur most frequently in the data set.

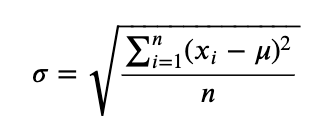
A data set can have more than one mode, and it can also have no mode.

## Standard Deviation

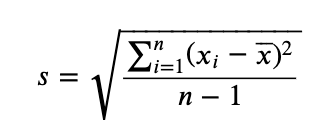
Standard deviation is a measure of dispersion of data values from the mean.

The formula for standard deviation is the square root of the sum of squared differences from the mean divided by the size of the data set.

For a Population



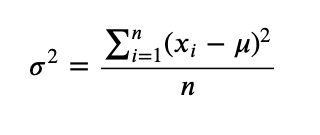
For a Sample



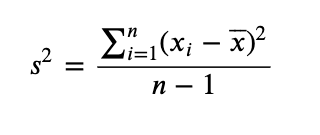
## Variance

Variance measures the dispersion of data from the mean. The formula for variance is the sum of squared differences from the mean divided by the size of the data set.

For a Population



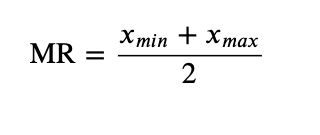
For a Sample



## Midrange

## 

The midrange of a data set is the average of the minimum and maximum values.



## Quartiles

Quartiles separate a data set into four sections.

The median is the second quartile Q2.

It divides the ordered data set into higher and lower halves.

The first quartile, Q1, is the median of the lower half not including Q2.

The third quartile, Q3, is the median of the higher half not including Q2.

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

## Interquartile Range

The range from Q1 to Q3 is the interquartile range (IQR).

𝐼𝑄𝑅=𝑄3−𝑄1

## Outliers

Potential outliers are values that lie above the Upper Fence or below the Lower Fence of the sample set.

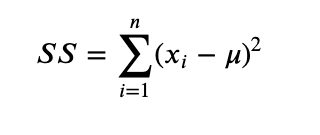
Upper Fence = 𝑄3+1.5×𝐼𝑄𝑅

Lower Fence = 𝑄1−1.5×𝐼𝑄𝑅

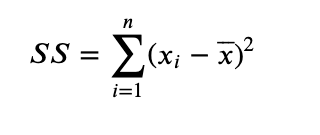
## Sum of Squares

The sum of squares is the sum of the squared differences between data values and the mean.

For a Population



For a Sample

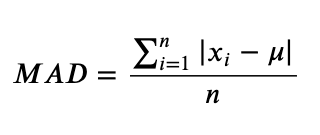


## 

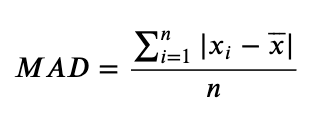
## Mean Absolute Deviation

Mean absolute deviation is the sum of the absolute value of the differences between data values and the mean, divided by the sample size.

For a Population



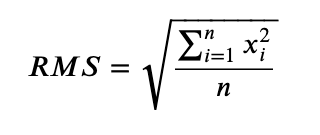
For a Sample



## Root Mean Square

The root mean square describes the magnitude of a set of numbers.

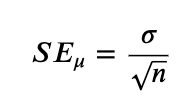
The formula for root mean square is the square root of the sum of the squared data values divided by n.



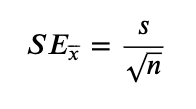
## Standard Error of the Mean

The standard error of the mean is calculated as the standard deviation divided by the square root of the count n.

For a Population



For a Sample



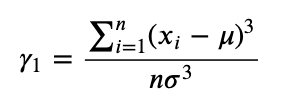
## Skewness

Skewness describes how far to the left or right a data set distribution is distorted from a symmetrical bell curve.

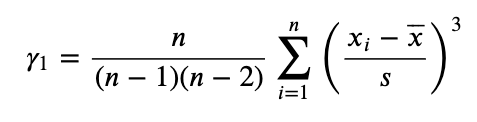
A distribution with a long left tail is left-skewed, or negatively-skewed.

A distribution with a long right tail is right-skewed, or positively-skewed.

For a Population



For a Sample



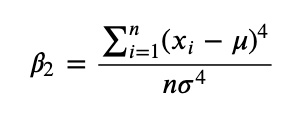
## Kurtosis

Kurtosis describes the extremeness of the tails of population distribution and is an indicator of data outliers.

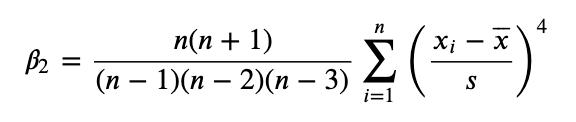
High kurtosis means that a data set has tail data that is more extreme than a normal distribution.

Low kurtosis means the tail data is less extreme than a normal distribution.

For a Population



For a Sample

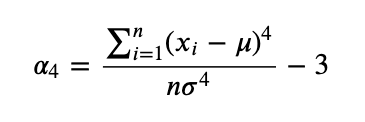


Kurtosis Excess

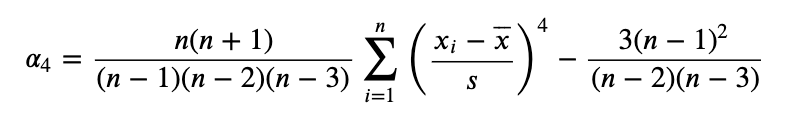
Excess kurtosis describes the height of the tails of a distribution rather than the extremity of the length of the tails.

Excess kurtosis means that the distribution has a high frequency of data outliers.

For a Population



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



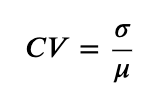
## Coefficient of Variation

The coefficient of variation describes the dispersion of data around the mean.

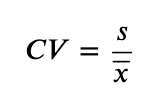
It is the ratio of the standard deviation to the mean.

The coefficient of variation is calculated as the standard deviation divided by the mean.

For a Population



For a Sample



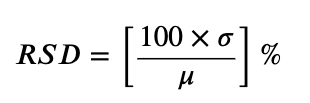
## Relative Standard Déviation

Relative standard deviation describes the variance of a subset of data from the mean.

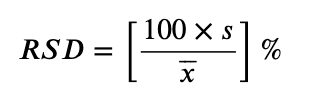
It is expressed as a percentage.

Relative standard deviation is calculated as the standard deviation times 100 divided by the mean.

For a Population



For a Sample



## Frequency

Frequency is the number of occurrences for each data value in the data set.

Frequency is used to find the mode of a data set.

# Reference list

*Descriptive statistics* (no date) *from Wolfram MathWorld*. Available at: https://mathworld.wolfram.com/topics/DescriptiveStatistics.html (Accessed: 26 March 2024).