­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 4](#_Toc162341959)

[Introduction 5](#_Toc162341960)

[Difference between Discrete Data and Continuous Data. 6](#_Toc162341961)

[Definition of discrete data. 6](#_Toc162341962)

[Definition of continuous data. 6](#_Toc162341963)

[Definition of Descriptive Statistics 6](#_Toc162341964)

[Common descriptive statistics in general include: 6](#_Toc162341965)

[Formulas and Calculations used on Descriptive Statistics 7](#_Toc162341966)

[Minimum 7](#_Toc162341967)

[Maximum 7](#_Toc162341968)

[Range 7](#_Toc162341969)

[Sum 8](#_Toc162341970)

[Size / Count 8](#_Toc162341971)

[Mean 8](#_Toc162341972)

[Median 8](#_Toc162341973)

[Mode 9](#_Toc162341974)

[Standard Deviation 9](#_Toc162341975)

[Variance 9](#_Toc162341976)

[Midrange 10](#_Toc162341977)

[Quartiles 10](#_Toc162341978)

[Interquartile Range 10](#_Toc162341979)

[Outliers 10](#_Toc162341980)

[Sum of Squares 11](#_Toc162341981)

[Mean Absolute Deviation 11](#_Toc162341982)

[Root Mean Square 11](#_Toc162341983)

[Standard Error of the Mean 12](#_Toc162341984)

[Skewness 12](#_Toc162341985)

[Kurtosis 13](#_Toc162341986)

[Kurtosis Excess 13](#_Toc162341987)

[Coefficient of Variation 14](#_Toc162341988)

[Relative Standard Déviation 14](#_Toc162341989)

[Frequency 15](#_Toc162341990)

# 

# Abstract

It is a useful and common practice to put the abstract in Times New Roman 12-point italics. Throughout this document, the styles used reflect the styles we suggest you use in your scientific report.

# Introduction

Format your submission using the following styles as a guide. If you are completing your report for a specific assignment, make sure you follow any formatting guidelines provided by your supervisor, lecturer or tutor.

# Difference between Discrete Data and Continuous Data.

## Definition of discrete data.

Discrete data is a type of quantitative data that includes no divisible figures and statistics you can count.

You typically write discrete data points as numbers that represent exact values.

You can often describe discrete data by using the phrase "the number of," such as the number of customers in a store.

Discrete data usually represents single events that have already occurred.

When reviewing discrete data, you can analyse exact figures like units sold on a specific day or the hours an employee worked during a certain week.

## Definition of continuous data.

Continuous data is a type of quantitative data that represents precise measurements of nearly any numeric value.

Often, a continuous data variable has many decimal points because it's an exact measurement between two defined points.

Measuring this type of data is common in industries requiring specific data, such as health care, manufacturing and research and development.

Continuous data may change over time, allowing businesses to analyse their operations and predict future trends.

For example, a business may track the amount of time it takes a team to complete projects.

# Definition of Descriptive Statistics

Descriptive statistics summarize certain aspects of a data set, population or sample, using numeric calculations.

## 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:

-

-

# Formulas and Calculations used on Descriptive Statistics

## Minimum

Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn

The minimum is the smallest value x1, on the left, starting from lowest to highest.

Min = 𝑥1 = min(𝑥𝑖)𝑛𝑖 = 1

## Maximum

The maximum is the largest value xn, on the right, starting from lowest to highest.

Ordering a data set:

x1 ≤ x2 ≤ x3 ≤ ... ≤ xn

Max=𝑥𝑛=max(𝑥𝑖)𝑛𝑖=1

Range

The difference between the minimum and maximum.

Range=𝑥𝑛−𝑥1

## Sum

The total of all data values.

x1 + x2 + x3 + ... + xn

Sum=∑𝑖=1𝑛𝑥𝑖

## Size / Count

The number of data points on the data set.

Size=𝑛=count(𝑥𝑖)𝑛𝑖=1

## Mean

Mean is the sum of all data divided by your size. Mean is also known as the “Average”.

For a Population

𝜇=∑𝑛𝑖=1𝑥𝑖𝑛

For a Sample

𝑥⎯⎯⎯=∑𝑛𝑖=1𝑥𝑖𝑛

## 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.

𝑝=𝑛+12

𝑥˜=𝑥𝑝

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

𝑝=𝑛2

𝑥˜=𝑥𝑝+𝑥𝑝+12

## 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

𝜎=∑𝑛𝑖=1(𝑥𝑖−𝜇)2𝑛‾‾‾‾‾‾‾‾‾‾‾‾‾‾√

For a Sample

𝑠=∑𝑛𝑖=1(𝑥𝑖−𝑥⎯⎯⎯)2𝑛−1‾‾‾‾‾‾‾‾‾‾‾‾‾‾√

## 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

𝜎2=∑𝑛𝑖=1(𝑥𝑖−𝜇)2𝑛

For a Sample

𝑠2=∑𝑛𝑖=1(𝑥𝑖−𝑥⎯⎯⎯)2𝑛−1

## Midrange

## 

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

MR=𝑥𝑚𝑖𝑛+𝑥𝑚𝑎𝑥2

## 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

𝑆𝑆=∑𝑖=1𝑛(𝑥𝑖−𝜇)2

For a Sample

𝑆𝑆=∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯)2

## 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

𝑀𝐴𝐷=∑𝑛𝑖=1|𝑥𝑖−𝜇|𝑛

For a Sample

𝑀𝐴𝐷=∑𝑛𝑖=1|𝑥𝑖−𝑥⎯⎯⎯|𝑛

## 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.

𝑅𝑀𝑆=∑𝑛𝑖=1𝑥2𝑖𝑛‾‾‾‾‾‾‾‾√

## 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

𝑆𝐸𝜇=𝜎𝑛√

For a Sample

𝑆𝐸𝑥⎯⎯⎯⎯=𝑠𝑛√

## 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

𝛾1=∑𝑛𝑖=1(𝑥𝑖−𝜇)3𝑛𝜎3

For a Sample

𝛾1=𝑛(𝑛−1)(𝑛−2)∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯𝑠)3

## 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

𝛽2=∑𝑛𝑖=1(𝑥𝑖−𝜇)4𝑛𝜎4

For a Sample

𝛽2=𝑛(𝑛+1)(𝑛−1)(𝑛−2)(𝑛−3)∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯𝑠)4

## 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

𝛼4=∑𝑛𝑖=1(𝑥𝑖−𝜇)4𝑛𝜎4−3

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

𝛼4=𝑛(𝑛+1)(𝑛−1)(𝑛−2)(𝑛−3)∑𝑖=1𝑛(𝑥𝑖−𝑥⎯⎯⎯𝑠)4−3(𝑛−1)2(𝑛−2)(𝑛−3)

## 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

𝐶𝑉=𝜎𝜇

For a Sample

𝐶𝑉=𝑠𝑥⎯⎯⎯

## 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

𝑅𝑆𝐷=[100×𝜎𝜇]%

For a Sample

𝑅𝑆𝐷=[100×𝑠𝑥⎯⎯⎯]%

## 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.