CANADIAN COLLEGE
OF TECHNOLOGY
AND BUSINESS

2023

Canadian College of Technology and Business

Invest in Yourself

Calculating Descriptive Statistics

CANADIAN COLLEGE
OF TECHNOLOGY
AND BUSINESS

Calculating Descriptive Statistics

Overview

- Previously, we represented categorical and numerical data through tables and graphs
- In this module, we will **summarize data** using numerical descriptive statistics

Calculating Descriptive Statistics

Central Tendency

- Describe the **central or average value** of a set of data
- The goal of central tendency is to find a single value that summarizes the center point of a dataset
- There are three main measures of central tendency
  - Mean
  - Median
  - Mode

Central Tendency

Mean

- The mean, or average, is calculated by adding up all the values in a dataset and dividing the sum by the number of values
- It's important to note that the mean is **sensitive to outliers** or extreme values in the data
  - If there are very large or very small values in the data, they can significantly impact the mean

Central Tendency

Mean

- Formula for the sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Study the formula

where:

$\bar{x}$ = The sample mean

x_i = The i^{th} value in the sample

n = The number of values in the sample

$\sum_{i=1}^{n} x_i$ = The sum of all values in the sample

- Sample mean is an estimate of population mean, which is mean of all values in entire population

Central Tendency

Mean

- Suppose we have a sample of 10 exam scores: 86, 92, 95, 87, 91, 89, 78, 84, 90, and 85

- The sample mean can be calculated as follows

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{86 + 92 + 95 + 87 + 91 + 89 + 78 + 84 + 90 + 85}{10} = 86$$

Apply the formula

- So, the sample mean of these exam scores is 86

Central Tendency

Mean

- It's important to note that the **sample mean is an estimate of the population mean**
  - If we had access to the scores of every student in the class, rather than just a sample of 10 students, we would calculate the population mean instead
  - The sample mean provides us with a rough idea of what the population mean might be, but it may not be the same as the population mean

Central Tendency

The Weighted Mean

- The weighted mean is a type of mean that **considers the relative importance or weight** of each data point
- It is used when different data points have different degrees of significance or importance in the analysis

Central Tendency

The Weighted Mean

- The weighted mean is calculated by multiplying each data point by its weight, summing up the products, and dividing by the total weight

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where:

$\bar{x}$ is the weighted mean

n is the total number of data points

x_i is the value of the i^{th} data point

w_i is the weight of the i^{th} data point

Central Tendency

The Weighted Mean

- Suppose we are analyzing the performance of a company that has three divisions: Division A, Division B, and Division C. We want to calculate the average revenue per employee across all three divisions, but want to **give more weight to the divisions that have more employees**

Division	Revenue	Employees
A	$5,000	400
B	$4,000	300
C	$3,000	300

$$\bar{x} = \frac{(400/1000) * 5000 + (300/1000) * 4000 + (300/1000) * 3000}{400/1000 + 300/1000 + 300/1000} = 4{,}100$$

Central Tendency

Median

- The median is the **middle value** in a dataset when the values are arranged in order
  - If there is an **odd number of values**, the median is simply the **middle value**
  - If there is an **even number of values**, the median is the **average of the two middle values**
- The median is a useful measure of central tendency, especially when the data set contains extreme values or outliers
- Unlike the mean, which can be heavily influenced by outliers, the median is not affected by extreme values

Central Tendency

Median

- For example, consider the following data set

$$\{3,1,9,5,7\}$$

- To find the median, we arrange the values in order

$$\{1,3,5,7,9\}$$

- Since there are 5 values, which is an odd number, the median is the middle value, which is 5

Central Tendency

Median

- Consider another example

$$\{12, 4, 10, 8, 6, 2\}$$

- Again, we arrange the values in order

$$\{2, 4, 6, 8, 10, 12\}$$

- This time, there are 6 values, which is an even number
  - So, the median is the average of the two middle values, which are 6 and 8
  - Therefore, the median is $\frac{(6+8)}{2} = 7$

Central Tendency

Mode

- The mode represents the **most frequently occurring value or values** in a dataset
- The mode is commonly used to describe the shape of a distribution
- In some cases, a data set may have multiple modes if there are multiple values that occur with the same frequency

Central Tendency

Mode

- Suppose we have a set of data that represents the number of pets owned by 20 individuals

  2, 3, 1, 4, 0, 2, 1, 1, 3, 0, 2, 1, 1, 1, 2, 2, 0, 3, 2, 3

- To find the mode of this data set, we need to identify the value that appears most frequently
- In this case, the number 1 appears the most often, occurring a total of 6 times
- Therefore, the mode of this data set is 1, which indicates that the most common number of pets owned by these individuals is 1

Central Tendency

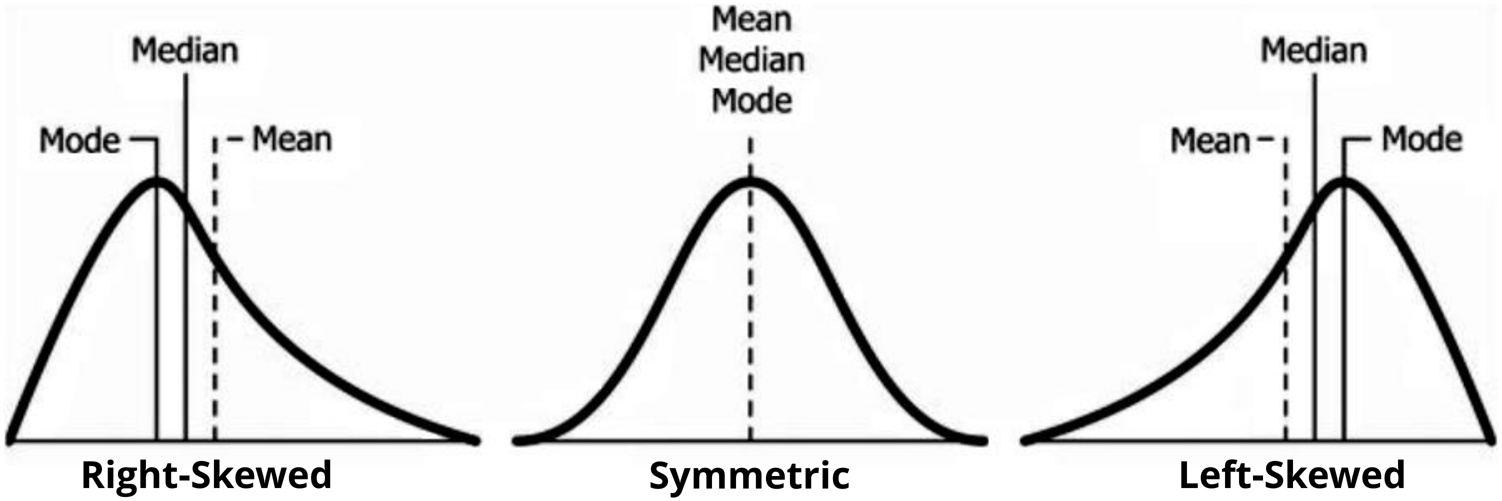
Mean, Median, and Mode

- Suppose we have the following set of data representing the number of hours slept by 10 individuals in a day: $\mathbf{6, 8, 7, 5, 6, 1, 8, 7, 7, 2}$
- Mean: $(6 + 8 + 7 + 5 + 6 + 1 + 8 + 7 + 7 + 2) / 10 = 57 / 10 = 5.7$
- Median:
  - Sort the values in ascending order: $1, 2, 5, 6, 6, 7, 7, 7, 8, 8$
  - Since there are 10 values in the data set, median is the average of the middle two values, which are 6 and 7. So, the median is: $(6 + 7) / 2 = 6.5$
- Mode: The value 7 appears the most often in the data set, occurring 3 times

Central Tendency

Shapes of Frequency Distributions

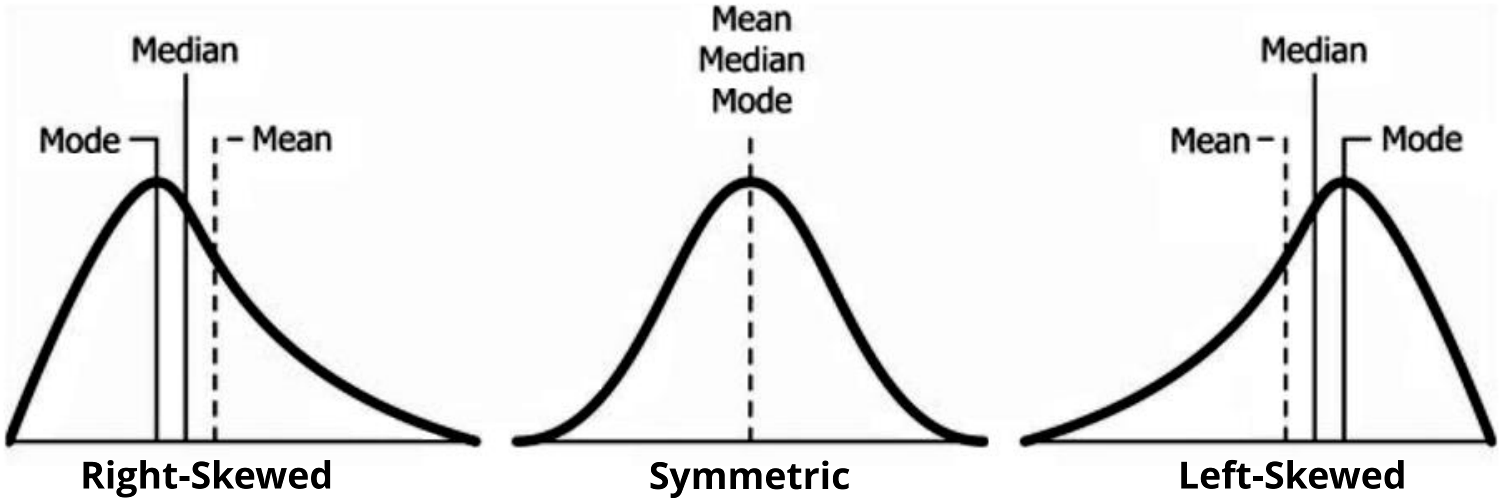
- Symmetric: A bell-shaped curve, with the **highest frequency occurring at the center** of the distribution and gradually decreasing toward either end
  - The **mean, median, and mode are all equal** in a perfectly symmetrical distribution



Central Tendency

Shapes of Frequency Distributions

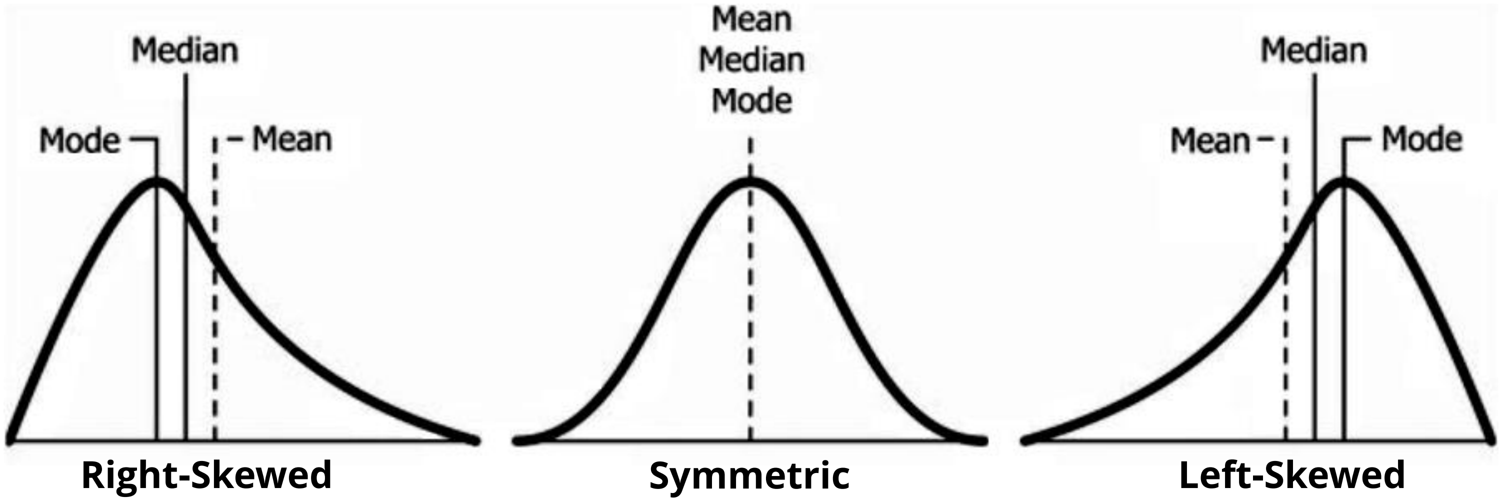
- Right-Skewed: A right-skewed distribution has a **long tail on the right side** of the distribution and a peak on the left side
  - The mean is greater than the median, which is greater than the mode
  - There are a few high values pulling the mean to the right of the median



Central Tendency

Shapes of Frequency Distributions

- Left-Skewed: A left-skewed distribution has a **long tail on the left side** of the distribution and a peak on the right side
  - The mean is less than the median, which is less than the mode
  - There are a few low values pulling the mean to the left of the median



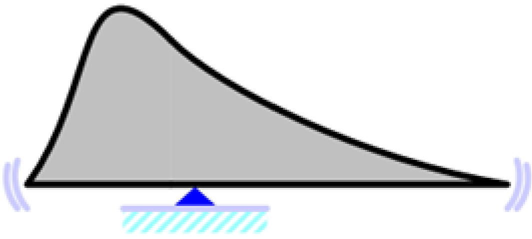
Central Tendency

Activity 1

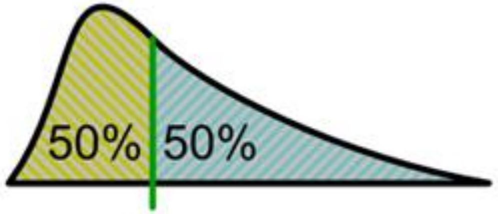
Central Tendency

The Choice of Measure of Central Tendency

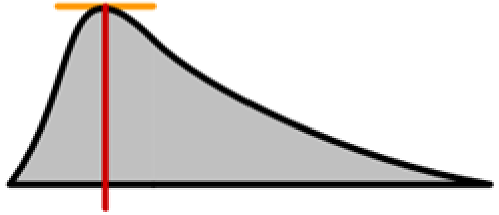
- Mean: The mean is a good measure of central tendency to use when the **dataset is symmetrical** and does not contain any extreme values (outliers)
- Median: The median is a better measure of central tendency to use when the **dataset is skewed** (i.e., not symmetrical) or contains extreme values (outliers)
- Mode: The mode is useful for **data consisting of nominal data** (i.e., data that cannot be ranked or have inherent order). It is also useful to know the most frequently occurring value in a dataset



Mean



Median



Mode

Measures of Variability

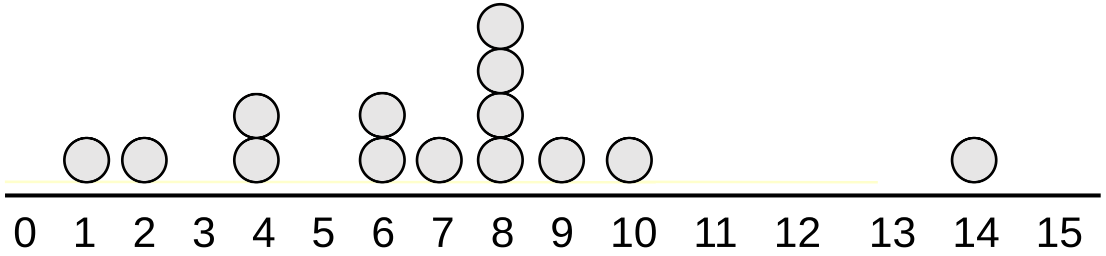
Overview

- Measures of variability, also known as measures of spread, are statistical tools used to describe the extent to which the values in **a dataset vary or are dispersed from the mean or median**
- There are several measures of variability, including
  - Range
  - Variance
  - Standard deviation
  - Interquartile range
  - Coefficient of variation

Measures of Variability

Range

- The range is the difference **between the largest and smallest values in a dataset**
- It is a simple and easy-to-calculate measure and provides a rough estimate of the spread of data
- However, it is **sensitive to outliers** and does not consider the distribution of the data between the smallest and largest values



Range = 14 - 1 = 13

Measures of Variability

Range

- Consider the following data set of daily high temperatures (in degrees Fahrenheit) for a week

  70, 72, 75, 69, 68, 74, 77

- To find the range of this data set, we first identify the largest and smallest values
  - Largest value: 77
  - Smallest value: 68
- Next, we subtract the smallest value from the largest value to obtain the range:

  77 - 68 = 9

- Therefore, the range of this data set is 9 degrees Fahrenheit. This means that the daily high temperatures vary by 9 degrees from the coldest day to the hottest day in the week

Measures of Variability

Variance

- The variance measures **how much individual data points in a data set vary from the mean**
- It is calculated as the average of the squared differences between each data point and the mean
- In statistics, there are two types of variance
  - Sample variance is used to estimate the variation of a sample
  - Population variance is used to measure the variation of a population

Measures of Variability

Sample Variance

- The formula for sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

where:

s^2 is the sample variance

n is the sample size

x_i represents each individual data point

$\bar{x}$ is the sample mean

Measures of Variability

Sample Variance

- The sample variance is often used to estimate the population variance, which is the measure of variability in a population
- To calculate the sample variance
  - Calculate sample mean $\bar{x}$ by summing all the data points and dividing by the sample size n
  - Calculate the difference between each data point and the sample mean $(x_i - \bar{x})$
  - Square each difference $(x_i - \bar{x})^2$ to eliminate any negative values
  - Add up all the squared differences $\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Divide the sum of squared differences by $n - 1$ to obtain the sample variance

Measures of Variability

Sample Variance

- Suppose we have the following dataset of 6 values

5, 8, 12, 6, 9, 10

1. Calculate the sample mean

$\bar{x}$ = (5 + 8 + 12 + 6 + 9 + 10) / 6 = 8.33

2. Calculate the difference between each data point and the sample mean

(5 - 8.33) = -3.33

(8 - 8.33) = -0.33

(12 - 8.33) = 3.67

(6 - 8.33) = -2.33

(9 - 8.33) = 0.67

(10 - 8.33) = 1.67

Measures of Variability

Sample Variance

3. Square each difference

$(-3.33)^2 = 11.09$

$(-0.33)^2 = 0.11$

$(3.67)^2 = 13.47$

$(-2.33)^2 = 5.43$

$(0.67)^2 = 0.45$

$(1.67)^2 = 2.78$

4. Add up all the squared differences

11.09 + 0.11 + 13.47 + 5.43 + 0.45 + 2.78 = 33.33

5. Divide the sum of squared differences by $n - 1$ to obtain the sample variance

s^2 = 33.33 / (6-1) = 6.666 ← The variance measures the variability/spread of the data points around the mean

Measures of Variability

Population Variance

- The population variance is a measure of the variability of a population
- The formula for population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

where:

σ^2 is the population variance

N is the population size

x_i represents each individual data point

μ is the population mean

Measures of Variability

Sample Variance and Population Variance

- Note that the only difference between the formula for population variance and sample variance is the denominator
- In the population variance formula, the denominator is N (the population size), while in the sample variance formula, the denominator is n-1 (the sample size minus 1)
- The use of n-1 instead of n in the denominator is known as Bessel's correction and is used to **adjust for the fact that sample variance tends to underestimate population variance**

Measures of Variability

Short-Cut Formulas for Variance

- There are shortcut formulas for the variance based on the sum of the squared values and the sum of the values
- The sample variance

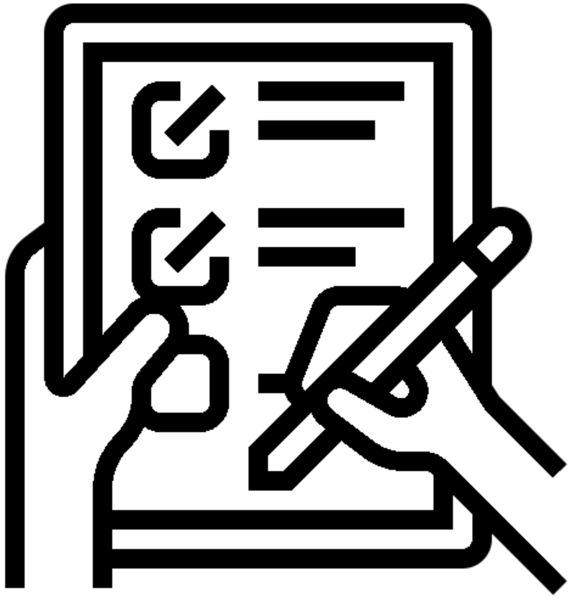
$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$$

- The population variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N}$$

Measures of Variability

Activity 2



Measures of Variability

Standard Deviation

- The standard deviation also **measures how far the data is from the mean but in a more intuitive unit** (i.e., the same unit as the original data)
- The sample standard deviation formula

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- The population standard deviation formula

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Measures of Variability

Standard Deviation

- Refer to the previous dataset of 6 values

  5, 8, 12, 6, 9, 10

- According to our calculation, the sample variance of this data set is 6.666
- We take the square root of the sample variance to get the sample standard deviation

$$s = \sqrt{s^2} = \sqrt{6.666} = 2.582$$

A measure of how far on average each data value is from the mean of the sample

Measures of Variability

Using the Mean and Standard Deviation

- The mean and standard deviation are two statistics often used together to describe a dataset
  - The mean provides a measure of the central tendency of the data
  - The standard deviation provides a measure of the spread or variability of the data

Using the Mean and Standard Deviation

The Coefficient of Variation

- A statistical measure used to **express the variability of a dataset relative to its mean**
- Mathematically, the coefficient of variation (CV) is calculated as follows
  - Sample coefficient of variation

$$CV = \frac{s}{\bar{x}}(100)$$

  - Population coefficient of variation

$$CV = \frac{\sigma}{\mu}(100)$$

Using the Mean and Standard Deviation

The Coefficient of Variation

- The CV is often used in situations where we want to **compare the variability of two or more datasets that have different means or units of measurement**
- A higher CV indicates a greater degree of variation relative to the mean, while a lower CV indicates less variation relative to the mean
  - A **smaller CV** indicates **more consistency** within a set of data values

Study it

Using the Mean and Standard Deviation

The Coefficient of Variation

- Suppose we want to compare the variability of the height of two different plant species
  - For species A, the mean height is 50 cm, and the standard deviation is 5 cm
  - For species B, the mean height is 175 cm, and the standard deviation is 10 cm
- At first glance, species B appears to have a greater variation in height than species A, but we need to consider the fact that the two species have different mean heights
- For two species, the CV would be

$CV_A = (5 / 50) \times 100 = 10\%$

$CV_B = (10 / 175) \times 100 = 5.71\%$

Although species B had a larger standard deviation it had the more consistent height

Using the Mean and Standard Deviation

The z-Score

- The z-score, a.k.a. the standard score, represents the **number of standard deviations** an observation or data point is from the mean of its distribution

- To calculate the z-score of a sample

$$z = \frac{x - \bar{x}}{s}$$

where: x is the data point we want to find the z-score for
$\bar{x}$ is the mean of the sample
s is the standard deviation of the sample

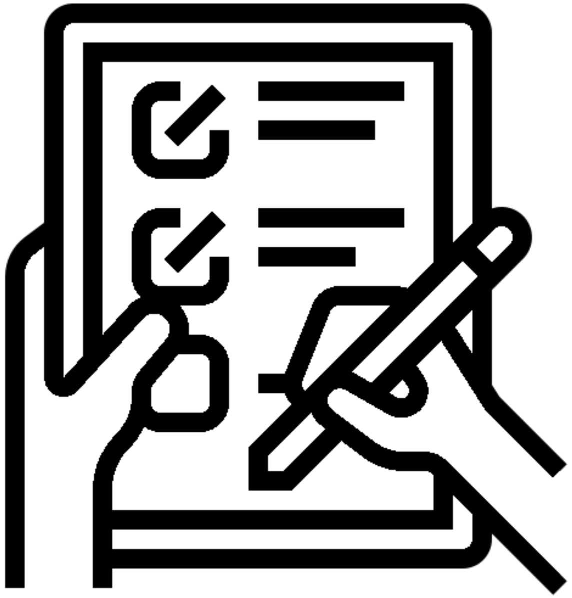
- To calculate the z-score of a population

$$z = \frac{x - \mu}{\sigma}$$

where: x is the data point we want to find the z-score for
μ is the mean of the population
σ is the standard deviation of the population

Using the Mean and Standard Deviation

Activity 3



Using the Mean and Standard Deviation

The z-Score

- A z-score is
  - 0 for values equal to the mean
  - positive for values above the mean
  - negative for values below the mean
- An **outlier** is a data point that is far from the mean, and it is identified as having a z-score greater than +3 or less than -3

Using the Mean and Standard Deviation

The z-Score

- Suppose we are a teacher grading students on a difficult exam
  - The class average is 75 with a standard deviation of 10
  - Assume a student receives a score of 95 on the exam. To determine how well he did relative to the rest of the class, we can calculate your z-score

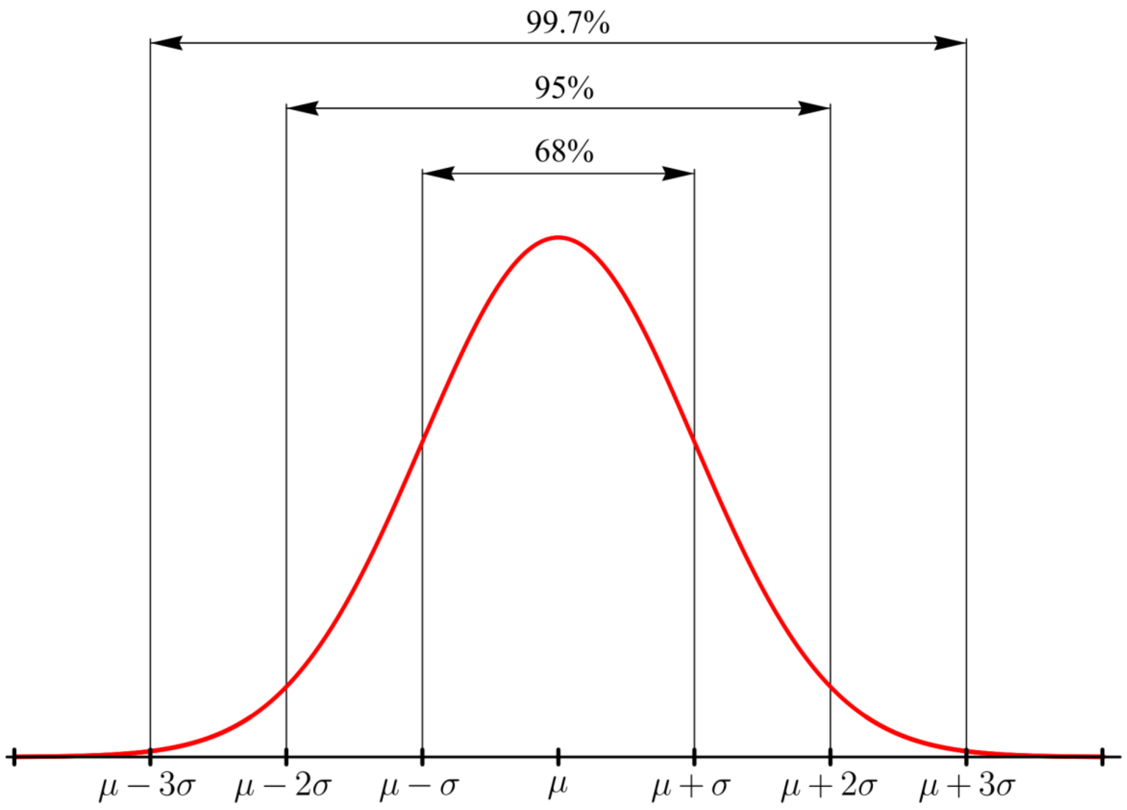
$$z = \frac{x - \mu}{\sigma} = \frac{95 - 75}{10} = 2$$

  - This means that his score is 2 standard deviations above the mean score of the class

Using the Mean and Standard Deviation

The Empirical Rule

- The empirical rule states that for a distribution follows a **bell-shaped, symmetrical curve centered around the mean**
  - About **68%** of the data falls within one standard deviation of the mean
  - About **95%** of the data falls within two standard deviations of the mean
  - About **99.7%** of the data falls within three standard deviations of the mean



Using the Mean and Standard Deviation

The Empirical Rule

- To express the z-score in terms of x
  - For a sample: $x \ = \bar{x} + zs$
  - For a population: $x \ = \mu + z\sigma$

Using the Mean and Standard Deviation

The Empirical Rule

- For a symmetric bell-shaped of salaries of a population with a mean of $50,000 and a standard deviation of $10,000. We want to find the salary range that contains about 95% of all earners
  - Approximately 95% of the values fall within two standard deviations of the mean
  - To find the interval containing about 95% of all the values, we can use the following formula

Interval = (mean – 2 * standard deviation, mean + 2 * standard deviation)

= (50000 – 2 * 10000, 50000 + 2 * 10000)

= (30000, 70000)

  - About 95% of the salaries will fall between $30,000 and $70,000

Using Probabilities to Make Decisions

Activity 4

Measures of Relative Position

Percentiles

- Percentiles are a way of describing the **relative position of a particular value in a dataset**
- It divides a dataset into 100 equal parts, with each percentile representing the percentage of data points that **fall below a certain value**
- Example
  - The 25th percentile represents the point at which 25% of the data falls below that value
  - The 75th percentile represents the point at which 75% of the data falls below that value

Measures of Relative Position

Percentiles

- To calculate a percentile, we first need to **sort the data from lowest to highest** and use the following formula to calculate the index point

$$i = \frac{p}{100} * n$$

where:

p is the percentile we are interested in

n is the total number of data points in the dataset

- If i is **not a whole number**, round i to the next whole number. The i^{th} position represents the value of interest

Study it

- If i is **a whole number**, the midpoint between the i^{th} and $(i+1)^{th}$ position is the value of interest

Measures of Relative Position

Percentiles

- Suppose we have the following dataset of 10 exam scores

$$67, 72, 75, 78, 80, 83, 85, 88, 90, 95$$

- To find the value at the 80th percentile

$$i = (80 / 100) * 10 = 8$$

- Since 8 is a whole number, take the average of the 8th and 9th data points, which are 88 and 90
- Value at 80th percentile = $(88 + 90) / 2 = 89$
- Therefore, the value at the 80th percentile in this dataset is 89

Measures of Relative Position

Quartiles

- Quartiles are values that **divide a dataset into four equal parts**
  - The first quartile (Q1) representing the 25th percentile
  - The second quartile (Q2) representing the 50th percentile (which is also the median)
  - The third quartile (Q3) representing the 75th percentile

Measures of Relative Position

Quartiles

- Suppose we have the following dataset of 10 exam scores

$$67, 72, 75, 78, 80, 83, 85, 88, 90, 95$$

- $Q1\ (25th)\colon\ i = (25/100) *\ 10 = 2.5$

  $\rightarrow \boldsymbol{Q1\ = 75}$

- $Q2\ (50th)\colon\ i = (50/100) *\ 10 = 5$

  $\rightarrow \boldsymbol{Q2\ =\ (80\ +\ 83)\,/\,2\ = 81.5}$

- $Q3\ (75th)\colon\ i = (75/100) *\ 10 = 7.5$

  $\rightarrow \boldsymbol{Q3\ = 88}$

Measures of Relative Position

Interquartile Range

- The interquartile range (IQR) defined as the difference **between the upper quartile (Q3) and the lower quartile (Q1)**

$$IQR = Q3 - Q1$$

Measures of Relative Position

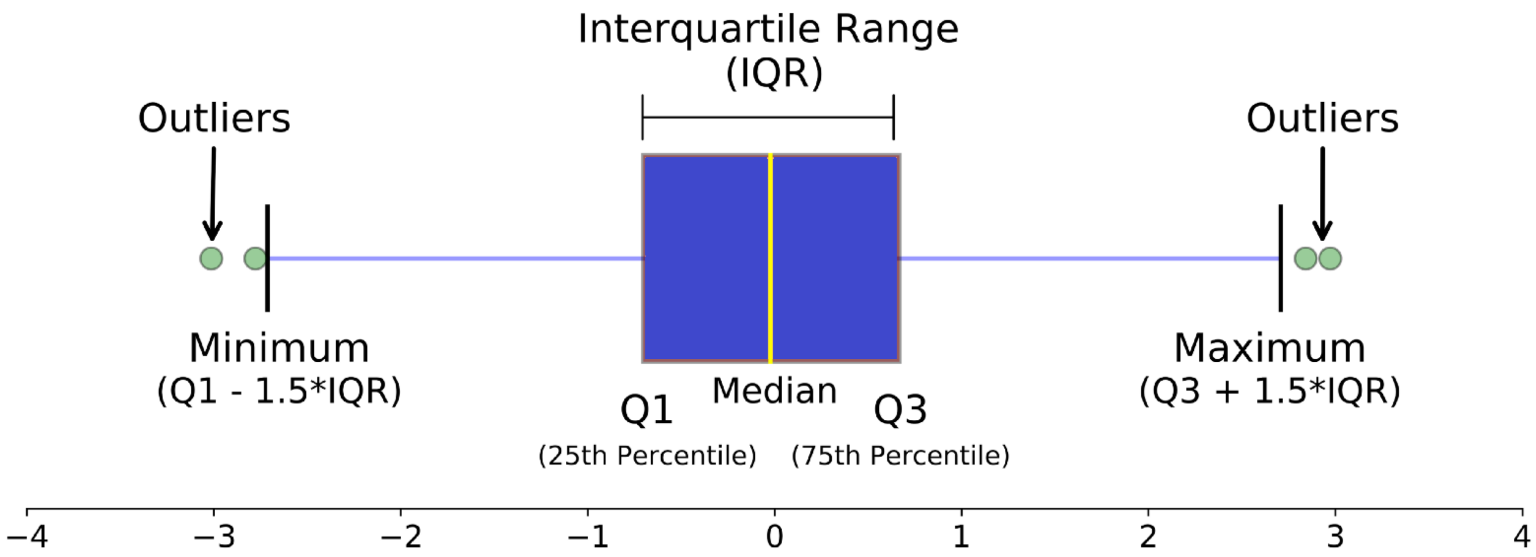
Box-and-Whisker Plots

- A box-and-whisker plot is a graphical representation of the distribution of a dataset that shows the median, quartiles, and extreme values of the data
- It is often used to identify potential outliers and compare the distribution of multiple datasets

Measures of Relative Position

Box-and-Whisker Plots

- It is particularly useful for **comparing the distributions** of multiple sets of data or **identifying outliers**
    - Outliers: Values that are significantly different from majority of the values in a dataset



Measures of Relative Position

Main Components

- **The box:** Represent the interquartile range (IQR) of the dataset
  - The IQR is the range of values that falls between the first quartile (25th percentile) and the third quartile (75th percentile) of the dataset
- **The median:** Represented as a line inside the box
  - The median is the value that separates the upper and lower halves of the dataset
- **The whiskers:** Extend from either end of the box
  - Represent the minimum and maximum values (excluding outliers)
- **The outliers:** Represented as individual points outside the whiskers

Measures of Relative Position

Box-and-Whisker Plots

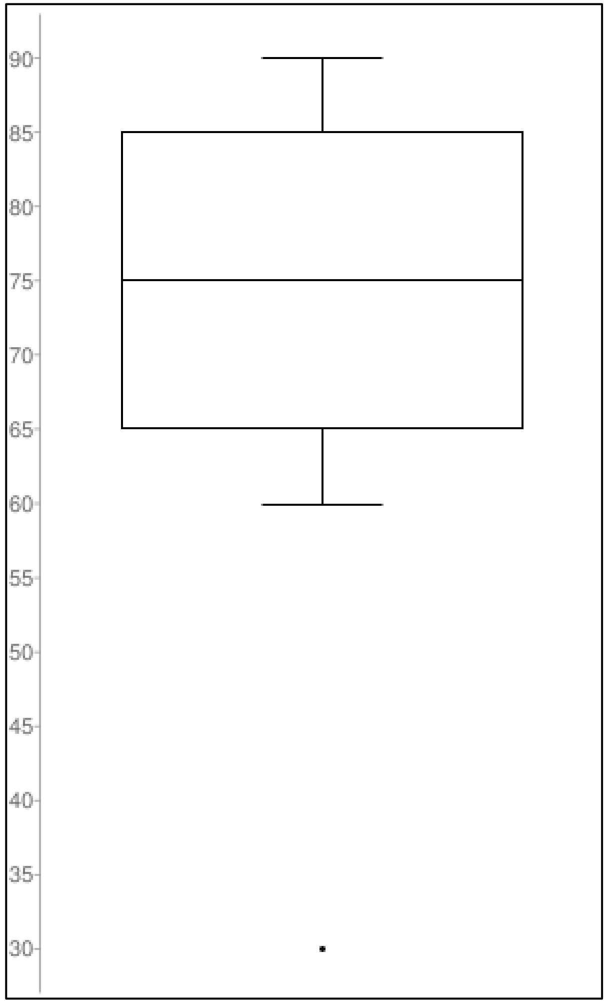
- Suppose we have the following dataset of exam scores

  30, 60, 65, 70, 72, 75, 78, 80, 85, 88, 90

- Population size: 11
- First quartile (Q1): 65
- Median (Q2): 75
- Third quartile (Q3): 85
- Interquartile Range (Q3 – Q1): 20
- Upper and lower limits
  - Upper Limit = Q3 + 1.5 (IQR) = 85 + 30 = 115
  - Lower Limit = Q1 – 1.5 (IQR) = 65 – 30 = 35
- Minimum: 30 (outlier) ⇒ 60 is the minimum
- Maximum: 90

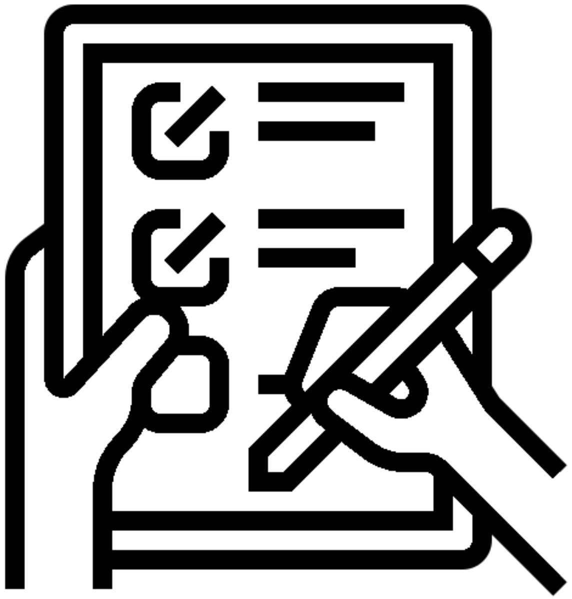
Actual upper calculated limit = max(data)

Actual lower calculated limit OR min(data)



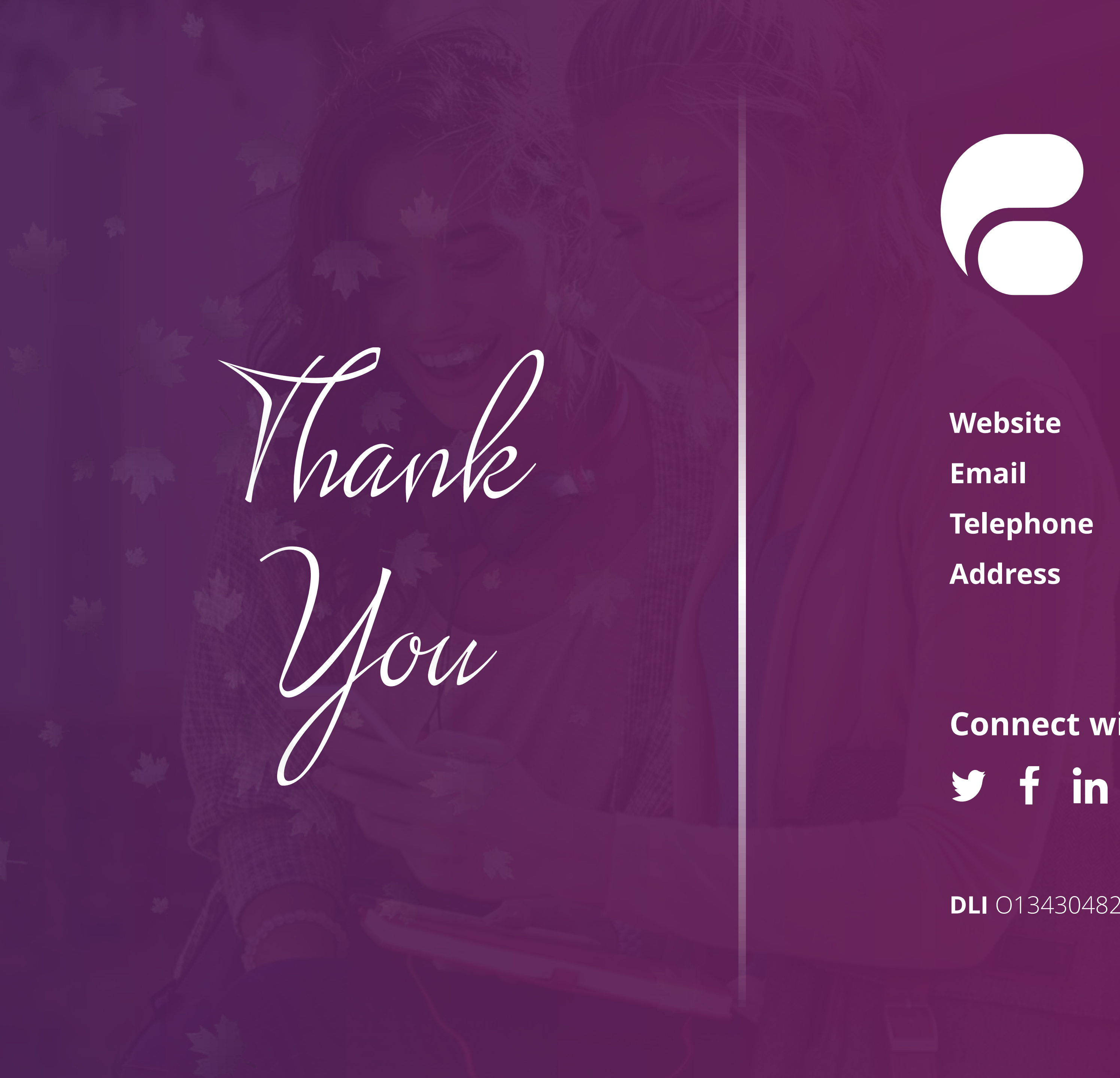
Measures of Relative Position

Activity 5



Any Questions?

CANADIAN COLLEGE
OF TECHNOLOGY
AND BUSINESS



Thank You

CANADIAN COLLEGE
OF TECHNOLOGY
AND BUSINESS

Website	www.canadianctb.ca
Email	info@canadianctb.ca
Telephone	+1 604-515-7880
Address	626 West Pender Street - Suite 600 Vancouver, British Columbia, V6B 1V9, Canada

Connect with CCTB

@CanadianCTB

DLI O134304821852