Week 4 - Descriptive Statistics in Mathematical Modelling

Introduction:

Welcome to this Module on Measures of Central Tendency and Measures of Variability in the context of mathematical modeling.

Descriptive statistics play a critical role in the context of mathematical modeling for several reasons. They provide a set of tools that help in summarizing and understanding the characteristics of a dataset, which is essential for the development, analysis, and validation of mathematical models. Here are some key reasons why descriptive statistics are useful in this context:

Understanding Data Distribution: Descriptive statistics help in understanding the distribution of the data. Measures such as mean, median, and mode give insights into the central tendency of the data, while variance and standard deviation offer information about the data's spread. Knowing the distribution helps in choosing the appropriate mathematical model that fits the data.

Identifying Outliers: Outliers can significantly affect the performance and accuracy of a mathematical model. Descriptive statistics enable the identification of outliers, which can then be analyzed to determine if they are errors or if they represent valuable extremes that should be included in the modeling process.

Data Summarization: In mathematical modeling, it is often necessary to present data in a summarized form to simplify analysis and interpretation. Descriptive statistics provide a concise description of the dataset through measures of central tendency and variability, making it easier to understand the overall characteristics of the data.

Model Assumption Verification: Many mathematical models make assumptions about the underlying distribution of the data (e.g., normality). Descriptive statistics can be used to verify these assumptions through graphical representations like histograms and Q-Q plots, ensuring that the chosen model is appropriate for the data.

Comparing Data Sets: Descriptive statistics are useful for comparing datasets, which is often required in modeling to understand changes over time or differences between groups. This comparison can inform model development and help in evaluating the model's effectiveness across different conditions or populations.

Feature Selection and Engineering: In the process of building a mathematical model, selecting the right features (variables) is crucial. Descriptive statistics can highlight relationships between variables through correlation coefficients and covariance, aiding in the selection of relevant features for the model.

Model Validation and Improvement: After a model has been developed, descriptive statistics are used to validate the model's performance by comparing predicted values against actual values. This process often involves analyzing residuals (the differences between observed and predicted values) to identify patterns that could suggest improvements to the model.

For the purposes of this model, we will be focusing on using descriptive statistics to understand data distributions, identify outliers and to summarize data.

Measures of Central Tendency

The idea of statistics is to boil down a set of numbers into something that can be easily digested by a person without the time, or the inclination to read all the numbers to come up to their own conclusions.

The Gini Index

The Gini Index measures how evenly wealth (or income) is shared within a country on a scale from zero to one. For example, if every individual had identical wealth within a country, the country would have a Gini Index of 0. However, if one individual had ALL the wealth within a country, the country would have a Gini Index of 1. The United States has a Gini Index score of .45. So what?

What we need is context. Remember I said that this was a tool for comparison, and therefore if we only have one data point it is kind of useless. So lets add context.

Country Gini	Index
Sweden	0.23
Canada	0.32
Brazil	0.54
South Africa	0.65
World	0.41

What conclusions can we reach about the US Gini Index score now we have this information?

We can also compare countries at different points of time

Example: The Gini index for the United States was .41 in 1997 and grew to .45 in 2007 and is now 39.8 as of 2021

We can also use visualisations to illustrate this (more on this next week)

Basic statistics you probably know

- Mean - Add them all the numbers up, and divide by how many there are
- Median - Put them in ascending order and pick the middle one.
- Mode - The most common value.

Measures of central tendency help us find the "center" or typical value of a dataset.

1. Mean (Average):

The mean is calculated by adding up all the values in a dataset and dividing by the number of observations. It is sensitive to outliers and can be skewed by extreme values.

The formula for the mean ( ) of a dataset with n observations is:

$ = (Σxi) / n $

2. Median:

The median is the middle value when the data is arranged in ascending or descending order. It is not affected by extreme outliers and is a robust measure of central tendency. For datasets with an even number of observations, the median is the average of the two middle values.

3. Mode:

The mode is the value that appears most frequently in the dataset. Some datasets may have multiple modes (bimodal, trimodal, etc.), while others may have no mode (no value occurs more than once).

Example:

Eleven students each did a test.

Their marks were as follows: 0, 2, 5, 5, 5, 7, 7, 8, 8, 9, 10

Calculate the mean, median and mode

Find the mean of grouped data

Often large amounts of data are grouped in a table. Here, the mean can be found using the following steps:

1. Multiply each value by its frequency
2. Add up all of the results from Step 1
3. Add up all of the frequencies
4. Divide the answer from Step 2 by the answer from step 3

Example:

The attendance figures for a 10-week course were as follows:

No. of times attended (x)	No. of students (f)
3	1
6	1
8	3
9	6
10	8

Example data

No. of times attended (x)	No. of students (f)	fx
3	1	3
6	1	6
8	3	24
9	6	54
10	8	80
Total	n = 19	$\sum fx = 167$

$$mean = \frac{\sum fx}{n} = \frac{167}{19} = 8.8$$

Find the median of grouped data

If the data are grouped, the median can be found as follows:

1. Work out the cumulative frequency (cf) for each value (ie. Add up the frequencies as you go along to obtain a running total)

2. Add 1 to the total frequency and then halve your answer. This tells you the position of the middle value
3. Find the lowest cf that is at least as large as the position found in Step 2
4. The value corresponding to the cf in Step 3 is the median

The median of our example data

No. of times attended (x)	No. of students (f)	Cumulative f
3	1	1
6	1	2
8	3	5
9	6	11
10	8	19

Following the steps this is (19+1) / 2 = 10th position, therefore the Median is 9

Find the mode of grouped data

No. of times attended	No. of students
3	1
6	1
8	3
9	6
10	8

Example where a response is a range of numbers

30 students were asked. 'How many times did you post something on Social Media over the weekend?'

1. 0
2. 1-5
3. 6-10
4. 11-15
5. 16+

Responses

No. of posts	No. of students
0	4
1-5	8
6-10	13
11-15	3
16+	2

How do we calculate averages here?

"1-5" is not a number, nor is "6-10" etc however, we could use the mid value for each and then find the averages.

No. of posts	No. of students	Mid-value
0	4	0
1-5	8	3
6-10	13	8
11-15	3	13
16+	2	18

- Mean = 6.77
- Median = 8
- Mode = 6-10

When do we use which average? (Nominal Variables)

- Categorical Data - Usually words
  - EG. England, USA, Mexico
  - We can't do calculation on words
  - The only average we can use is the Mode
- Ranked Categories (Eg, Small, Medium, Large)
  - We could use the mode
  - Is is more useful to find the median, we do this by assigning number values to each category (usually starting with 1 as the lowest rank)
- Scale Variables
  - Counts or measures should use a recognisable unit
  - Can use Mean, Median or Mode
  - The selection is based on the appropriateness of the measure

Consider a survey in which the question asked is 'How much do you like vodka?' | Response | Not at all | A little | Quite like it | Love it | |———-|———|———-|———|———| | Number Values | 1 | 2 | 3 | 4 | | Frequency | 5 | 3 | 9 | 2 | | Cumulative Frequency | 5 | 8 | 17 | 19 |

In this case. The total number is 19. The mean score is ? The median score is ? The modal score is ?

When shoudn't we use the mean? (Example 1)

Imagine a bar in Seattle, Washington. Here you have 5 people who, coincidentally, earn exactly $35,000 a year. We can say, correctly, that the average salary of this group of people is $35,000. What if, Bill Gates walks into the bar and joins this group of people?

- Bill Gates currently has an annual income of $12,000,000,000.
- Calculate the average income of this group
- Average salary of this group is now - $2,000,029,167
- Is this appropriate? Is there a better method?

When shouldn't we use the mean? (Example 2)

How many arms does a human have on average? The vast majority of people have 2 arms, but some will have 1 arm or no arms. Hence the mean may be around say 1.995. However, saying a human has 1.995 arms on average could be considered misleading, as the huge majority have more than that. It would probably be more useful to state the mode which is 2 arms

Measures of Variance

We talked about the idea of using Measures of Central Tendency (Averages), as a means of describing and summarising what data looks like. However, that's not the whole picture. In this seminar we will look at measures of variance as an additional tool for describing and summarising data.

Example

2 groups of people were asked 'How many hours per week do you exercise?'

Group A gave the following responses 4, 5, 3, 4, 5, 3, 4 Group B gave the following responses 0, 3, 10, 1, 7, 3, 4

The mean for group A was $\frac{4+5+3+4+5+3+4}{7} = 4$

The mean for group B was $\frac{0+3+10+1+7+3+4}{7} = 4$

Both groups have the same average But the values for group B are more spread out. They vary more.

What does this mean? Not only do we need to find a way to find the central point of a dataset (Mean, median or mode) but we also need to find a way to measure the spread or variability of a dataset.

There are several measures for variability (or spread). We are going to look at 4.

- Range
- Inter-quartile Range
- Variance
- Standard Deviation

Range

> **Range formula**
>
> $Range = Highestvalue - LowestValue$

Range formula

For Gp A: 4, 5, 3, 4, 5, 3, 4

$Range = 5 - 3 = 2$

For Gp B: 0, 3, 10, 1, 7, 3, 4

$Range = 10 - 0 = 10$

Interquartile Range

Put them in ascending order and calculate the value 3/4 of way – the value ¼ of the way

> **IQR formula**
>
> $IQR = upperquartile - lowerquartile = Q3 - Q1$
> Q1 is $\frac{n+1}{4}^{th}$ value, Q3 is $3(\frac{n+1}{4})^{th}$ value

Example Gp A: 3, 3, 4, 4, 4, 5, 5

Q1 is $\frac{7+1}{4}^{th}$ value = 2nd value = 3

Q3 is $3(\frac{7+1}{4})^{th}$ value = 6th value = 5

Therefore $IQR = Q3 - Q1 = 5 - 3 = 2$

Calculate this for Group B. Gp B: 0, 1, 3, 3, 4, 7, 10

Example Gp B: 0, 1, 3, 3, 4, 7, 10

Q1 is $\frac{7+1}{4}^{th}$ value = 2nd value = 1

Q3 is $3(\frac{7+1}{4})^{th}$ value = 6th value = 6

Therefore $IQR = Q3 - Q1 = 7 - 1 = 6$

The IQR is usually preferable to the range because it is unaffected by outliers (extreme values)

If Gp A responses were: 3, 3, 4, 4, 4, 5, 5 then range = 2 and IQR = 2 however, If the highest response had been 50 hrs Gp A: 3, 3, 4, 4, 4, 5, 50 then Range = 47, IQR = 2. In this situation, IQR is better because 50 might have been a typo or a lie. Even if the 50 was true, the IQR is more representative of the whole group

Variance formula

$Variance = \frac{\sum(x-\bar{x})^2}{n-1}$

Variance For 2 groups, calculate the variance. Gp A: 4, 5, 3, 4, 5, 3, 4

- $\sum x^2 = 4^2 + 5^2 + 3^2 + 4^2 + 5^2 + 3^2 + 4^2 = 116$
- $\bar{x} = 4$
- $var = \frac{116-7(4^2)}{7-1} = 0.667$

Variance Gp B: 0, 3, 10, 1, 7, 3, 4 - $\sum x^2 = 0^2 + 3^2 + 10^2 + 1^2 + 7^2 + 3^2 + 4^2 = 184$ - $\bar{x} = 4$ - $var = \frac{184-7(4^2)}{7-1} = 12$

Standard Deviation formula

$sd = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

Using the variance calculations Gp A: 4, 5, 3, 4, 5, 3, 4

$\sqrt{\frac{116-7(4^2)}{7-1}} = 0.816$

Gp B: 0, 3, 10, 1, 7, 3, 4

$\sqrt{\frac{184-7(4^2)}{7-1}} = 3.46$

Interpretation of standard deviation - It roughly tells you how much the data vary from the mean on average

Which measure of variability should we use?

The standard deviation is usually best to use, and is convention. However if

a. there are some outliers OR
b. it is an ordinal variable (one which has ordered categories)

then you should use the IQR

Applications in Mathematical Modeling:

Measures of central tendency and variability are vital in mathematical modeling for various reasons:

- They help summarize and understand the data used to build models.
- They provide insights into the distribution and characteristics of the data.
- They help identify outliers and understand their impact on the model.
- They guide the selection of appropriate probability distributions for modeling uncertainty.

Conclusion:

Measures of central tendency (mean, median, mode) and variability (range, variance, standard deviation, IQR) are essential statistical tools for summarizing and understanding data in the context of mathematical modeling.

These measures assist in characterizing the central behavior and spread of data, aiding model development, and decision-making.