Exploratory Data Analysis

Madiba Hudson-Quansah

Contents

Chapter 1

Introduction

Definition 1.0.1: Data

Pieces of information about individuals (specific person / object) organized into variables (specific characteristic of the individual).

Definition 1.0.2: Dataset

A set of data identified with particular circumstances.

Variables can be classified into one of two types:

- Categorical / Qualitative
- Quantitative

Definition 1.0.3: Categorical / Quantitative Variable

Take category or label values and places an individual into one of several mutually exclusive groups.

Definition 1.0.4: Quantitative Variable

Take numerical values and represents a measurement.

Chapter 2

Module 4: Examining Distributions

Definition 2.0.1: Distribution

The distribution of a variable tells us what values the variable takes and how often it takes these values.

2.1 Distribution of One Categorical Variable

In order to summarize the distribution of a categorical variable we must first create a table of all the values the category can take (categories) and how often it takes these values (counts). This is called a **frequency table**.

In order to visualize the summaries we've made we can use one of two graphs:

- Pie Chart
- Bar Graph

2.2 Distribution of One Quantitative Variable

To visualize a summary of a quantitative variable we can use one of three graphs:

- Histogram
- Stem plot
- Box plot

2.2.1 Histogram

To construct a histogram we must first divide the range of our data into equal sized intervals called **bins** / **classes**. We then count how many observation fall into each bin and construct a frequency table with the bins and counts. We then plot the bin on the x-axis and the count on the y-axis.

Definition 2.2.1: Relative Frequency

The proportion of observations in a category. Calculated by dividing the count of observations in a category by the total number of observations.

In interpreting the histogram we must consider the following features of the distribution:

- Shape
- Centre
- Spread

- Outliers

Where the first three describe the distribution as a whole and the patterns found within and the last highlights deviations from that pattern.

2.2.1.1 Shape

The shape of a distribution is described by its **modality / peakedness** and **symmetry/skewness**.

Definition 2.2.2: Modality

Number of peaks (modes) in a distribution.

Definition 2.2.3: Symmetric Distribution

A distribution is symmetric if the right and left sides of the histogram are approximately mirror images of each other.

Definition 2.2.4: Skewed Distribution

A distribution is skewed if one of its tails is longer than the other.
- **Right Skewed** if the right tail is longer.
- **Left Skewed** if the left tail is longer.

Definition 2.2.5: Unimodal

A distribution with one mode, i.e. a central peak the observations are concentrated around.

Definition 2.2.6: Bimodal

A distribution with two modes, i.e. two central peaks the observations are concentrated around.

Definition 2.2.7: Multimodal

A distribution with more than two modes, i.e. more than two central peaks the observations are concentrated around.

Definition 2.2.8: Uniform

A distribution with no mode, i.e. the observations are evenly distributed across the range of the data.

2.2.1.2 Centre

The value that divides the distribution so that approximately half the observations take smaller values, and half take larger values.

2.2.1.3 Spread

The approximate range covered by the data. i.e. The smallest value to largest value.

2.2.1.4 Outliers

Observations that fall outside the overall pattern of the distribution.

2.2.2 Stem Plot / Stem and Leaf Plot

To construct a stem plot we must

- Separate each observation into a **stem** and a **leaf**. The stem being everything except the right post digit and the leaf being the right most digit, e.g. 123 would have a stem of 12 and a leaf of 3.
- Write the stems in a vertical column in ascending order.
- Go through the data points and match each leaf to its stem in ascending order.

Question 1

Construct a stem plot for the following Dataset
34 34 27 37 42 41 36 32 41 33 31 74 33 49 38 61 21 41 26 80 42 29 33 36 45 49 39 34 26 25 33 35 35 28 30 29 61 32 33 45 29 62 22 44

Solution:

2	1 2 5 6 6 7 8 9 9
3	0 1 2 2 3 3 3 3 3 4 4 5 5 6 6 7 8 9
4	1 1 1 2 2 4 4 5 9 9
5	
6	1 1 2
7	4
8	0

2	1 2
2	5 6 6 7 8 9 9
3	0 1 2 2 3 3 3 3 3 4 4
3	5 5 6 6 7 8 9
4	1 1 1 2 2 4 4
4	5 9 9
5	
5	
6	1 1 2
6	
7	4
7	
8	0

When skewed right the stem plot can be visualize to identify the skewness.

Note:-

The advantages of the stem plot are:
- It preserves the original data.
- It sorts the data.
- It is easy to construct for small datasets.

2.3 Numerical Measures

To get a more accurate description of a discrete quantitative variable we can use numerical measures.

2.3.1 Measures of Centre

The numerical measure of a distribution's centre is basically telling us what is a typical value for the variable. The three main numerical measures of centre are the **mean**, **median** and **mode**.

2.3.1.1 Mode

The most common occurring value in a distribution.

2.3.1.2 Mean

The average of a set of observations i.e. the sum of all the observations divided by the number of observations. i.e.

$$\bar{x} = \frac{\sum f}{n}$$

Where f is the frequency of the observation and n is the number of observations.

2.3.1.3 Median

The midpoint of a distribution. i.e. the value that divides the distribution so that approximately half the observations take smaller values, and half take larger values.

It can be found by:

- First Ordering the observations in ascending order.
- Consider whether n (number of observations) is even or odd. If:
  - Even - The median is the average of the two middle observations, i.e.

    $$\frac{M_1 + M_2}{2}$$

    where M_1 and M_2 are the two middle observations, and where M_1 and M_2 found at the $\frac{n}{2}$ and $\frac{n}{2} + 1$ positions respectively.
  - Odd - The median is the middle observation. i.e. M where M is found at the $\frac{n+1}{2}$ position

2.3.1.4 Comparing the Mean and Median

The mean is being the average of all the observations is more sensitive to outliers than the median. This means that for:

- A symmetric distribution with no outliers - $\overline{x}$ (the mean) $\approx$ (approximately equal) M (the median).
- A right skewed distribution and/or datasets with high outliers - $\overline{x} > M$
- A left skewed distribution and/or datasets with low outliers - $\overline{x} < M$

Therefore it is best to use mean ($\overline{x}$) as a measure of centre for symmetrical distributions with no outliers, Otherwise median (M) is a better measure of centre.

2.3.2 Measures of Spread

The numerical measure of a distribution's spread is basically telling us how spread out the data is. The three main numerical measures of spread are the **range**, **interquartile range** and **standard deviation**.

2.3.2.1 Range

The range of a distribution is the difference between the largest and smallest observations. i.e. $R = x_{max} - x_{min}$

2.3.2.2 Interquartile Range

The interquartile range (IQR) measures variability of a distribution by giving us the range covered by the middle 50% of the observations instead of the whole range covered by all the observations. To calculate the IQR we must

- Arrange the observations in ascending order.
- Find the median.
- Find the median of the lower 50% of observations also called the first/lower quartile (Q1). We can find the position of the lower quartile using the formula:

$$Q1_{th} = \frac{1}{4}(n + 1)$$

- Find the median of the upper 50% of observations also called the third/upper quartile (Q3). We can find the position of the upper quartile using the formula:

$$Q3_{th} = \frac{3}{4}(n+1)$$

- Calculate the IQR by subtracting Q1 from Q3. i.e. $IQR = Q3 - Q1$

2.3.2.2.1 Using the IQR to Identify Outliers. The IQR is used as the basis for a rule of thumb for identifying outliers called the **1.5 × IQR Rule**. According to this rule, observations that fall more than 1.5 × IQR above the third quartile or below the first quartile are considered outliers.

2.3.2.2.2 Box and Whisker / Box Plot. A box plot is a graphical display of the five number summary. It is constructed by drawing a box with the first side at the lower quartile and the second side at the upper quartile, then drawing a line through the box at the median. Finally we draw vertical lines at the maximum and minimum values and connect them to the box with horizontal lines.

Definition 2.3.1: Five Number Summary

The combination of the minimum, lower quartile, median, upper quartile and maximum of a distribution.

Another way of constructing box plots is the **Tukey method**. This consists of:

- Calculating the IQR
- Finding 1.5 × IQR and adding it to the upper quartile. If the value is greater than or equal to the maximum value of the distribution then the maximum value is used as the upper whisker. Otherwise the largest value in the distribution that is less than the upper quartile + 1.5 × IQR is used as the upper whisker.
- Subtracting 1.5 × IQR from the lower quartile. If the value is less than or equal to the minimum value of the distribution then the minimum value is used as the lower whisker. Otherwise the smallest value in the distribution that is greater than the lower quartile − 1.5 × IQR is used as the lower whisker.
- If the maximum or minimum values are not used as the whiskers of the boxplot they are instead denoted by asterisks (*).

2.3.2.3 Standard Deviation

The standard deviation measures how far the observations are from their mean $\overline{x}$ .i.e. the average distance between a data point and the mean. The standard deviation of a sample is denoted by s and is calculated by:

$$s = \sqrt{\frac{\sum (X - \overline{x})^2}{n-1}}$$

Where X is the observation, $\overline{x}$ is the mean of the observations and n is the number of observations.

Note:-

The standard deviation is always positive.

Note:-

Since the SD is dependent on the mean, it should be used as a measure of spread only when the mean is used as a measure of centre. Also due to the SD's dependence on the mean it is also sensitive to outliers.

Theorem 2.3.1 Standard Deviation Rule / Empirical Rule

For a symmetric bell shaped distribution, i.e. normal distribution, the following rule applies:

- Approximately 68% of the observations fall within 1 SD of the mean .i.e $68\% \times n$ lies between $\overline{x} \pm s$
- Approximately 95% of the observations fall within 2 SD of the mean .i.e. $95\% \times n$ lies between $\overline{x} \pm 2s$
- Approximately 99.7% of the observations fall within 3 SD of the mean. .i.e. $99.7\% \times n$ lies between $\overline{x} \pm 3s$

Note:-

When comparing two distributions with different units of measurement using standard deviation, we must use the **Coefficient of Variation** for each distribution. This is calculate by:

$$CV = \frac{S}{\overline{x}} \times 100$$

Where S is the standard deviation and $\overline{x}$ is the mean.

2.3.3 Choosing Numerical Measures

Use $\overline{x}$ and s as measures of centre and spread only for reasonably symmetric distributions with no outliers. Otherwise use M and IQR.

Chapter 3

Module 5: Examining Relationships

Definition 3.0.1: Independent / Explanatory Variable

The variable that claims to explain, predict or affect the response variable.

Definition 3.0.2: Dependent / Response Variable

The variable that measures an outcome or result of a study.

Further classification of variables:

- Categorical Explanatory / Quantitative response, $C \to Q$
- Categorical Explanatory / Categorical response, $C \to C$
- Quantitative Explanatory / Quantitative response, $Q \to Q$
- Quantitative Explanatory / Categorical response, $Q \to C$

When confronted with a research question that involves exploiting the relationship between two variables, the first step should be determining which of the four cases above applies to the variables in question. This will help us determine what statistical methods to use to answer the question.

3.1 Case $C \to Q$

In this case we essentially compare the distributions of the quantitative response variable for each category of the explanatory variable. This can be done using side by side **Boxplots** and **Descriptive Statistics / Five Number Summary**.

3.2 Case $C \to C$

In this case we essentially compare the distributions of the categorical response variable for each category of the explanatory variable. This can be done using a **two way table** comparing the counts of each category of the response variable for each category of the explanatory variable, and finding the proportions of each category of the response variable for each category of the explanatory variable.

3.3 Case $Q \to Q$

In this case we examine the relationship between the two quantitative variables by plotting the explanatory variable on the x-axis and the response variable on the y-axis. We then look for patterns in a **Scatter plot**.

3.3.1 Scatter Plot

3.3.1.1 Interpreting the Scatter plot

When interpreting a scatter plot we must consider the following features of the distribution:

- Direction
- Form
- Strength
- Outliers

3.3.1.1.1 Direction The direction of a scatter plot can be classified as either **positive**, **negative** or **no relationship**.

Definition 3.3.1: Positive relationship

An increase in one of the variables is associated with an increase in the other.

Definition 3.3.2: Negative Relationship

An increase in one of the variables is associated with a decrease in the other.

Definition 3.3.3: No Relationship

There is no apparent relationship between the two variables.

3.3.1.1.2 Form The general shape of the relationship. Some common shapes **Linear**, **Curvilinear** and **Clusters**.

Definition 3.3.4: Linear

Looks like points scattered around a straight line.

Definition 3.3.5: Curvilinear

Looks like points scattered around a curved line.

Definition 3.3.6: Clusters

Looks like points gathered around a particular point.

3.3.1.1.3 Strength How closely the data follows the form of the relationship. A relationship's strength can be classified as either **Strong**, **Moderate** or **Weak**.

Definition 3.3.7: Strong

The points follow the form of the relationship very closely.

Definition 3.3.8: Moderate

The points follow the form of the relationship moderately closely.

Definition 3.3.9: Weak

The points follow the form of the relationship weakly.

3.3.1.1.4 Outliers

Data points that deviate from the pattern of the relationship.

3.3.2 Linear Relationships

3.3.2.1 The Correlation Coefficient (r)

Definition 3.3.10: Correlation Coefficient (r)

A numerical measure of the strength and direction of a linear relationship between two quantitative variables.

It is calculated by:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_x} \right) - \left(\frac{y_i - \overline{y}}{S_y} \right)$$

Where x_i and y_i are the ith observations of the explanatory and response variables respectively, S_x and r can take values between -1 and 1. Where:

- $r > 0$ - Positive relationship
- $r < 0$ - Negative relationship
- $r = 0$ - No relationship

The closer r is to 1 or -1 the stronger the relationship. The closer r is to 0 the weaker the relationship.

As a numerical measure r has several properties to take note of:

- r is not dependent on the units of measurement of the variables.
- r only measures the strength of a linear relationship, so it is not appropriate for non-linear relationships as it tries to fit a straight line to a non-linear relationship.
- r by itself is not sufficient to determine the form of a relationship between two variables.
- r is strongly affected by outliers.

3.3.2.2 Linear Regression

Definition 3.3.11: Regression

The technique that specifies the dependence of the response / dependent variable on the explanatory / independent variable.

Definition 3.3.12: Linear Regression / Line of Best Fit

Regression in the form of a linear function.

In constructing the line of best fit there are may methods that can be used. The most common method is the **Least Squares Method**.

Definition 3.3.13: Least Square Method / Criterion

The method of finding the line of best fit by minimizing the sum of squared vertical deviations of the data points from the line.

The resulting line of best fit is called the **Least-Squares Regression Line**. As with all straight lines it's equation is of the form

$$Y = a + bX$$

Where Y is the response variable, X is the explanatory variable, a is the y-intercept and b is the slope of the line. To calculate the y-intercept and slope we need the following:

- $\overline{X}$ - Mean of the independent variables
- S_X - Standard deviation of the independent variables
- $\overline{Y}$ - Mean of the dependent variables
- S_Y - Standard deviation of the dependent variables
- r - Correlation coefficient.

Given these values we can thus find a and b using the following formulas:

$$b = r\left(\frac{S_Y}{S_X}\right)$$
$$a = \overline{Y} - b\overline{X}$$

3.4 Causation and Lurking Variables

Association between two variables does not imply causation. There may be lurking variables that may be responsible for the observed relationship between the two variables.

Definition 3.4.1: Lurking Variable

A variable that is not among the independent or dependent variables in a study by could substantially influence the interpretation of the relationship among those variables.