Exploratory Data Analysis

Jerry Kiely

2023-11-27

Contents

1 Data Management

1.1 Chapter 1

To be completed...

2 Descriptive Statistics

2.1 Measures of Central Tendency and Variation

2.1.1 Sources and Types of Data

2.1.1.1 Sources of Data

1. Primary Data

- the data is collected by the investigator himself / herself for a specific purpose
- direct method of data collection
- eg. data collected for research through questionnaires, interviews

2. Secondary Data

- the data is collected by someone else, but being used by the investigator for some other purpose
- an indirect method of data collection
- eg. census data being used to study the impact of education on income

2.1.1.2 Types of Data

1. Structured Data

- information is stored with a high degree of organization
- contains qualitative data, quantitative data, or a mixture of both
- eg. data arranged in an excel file, in rows and columns

2. Unstructured Data

- information that either does not have a pre-defined data model and / or is not organized in a pre-defined manner
- eq. emails, tweets, blogs, etc.

2.1.2 Measurement Scales

1. Nominal Scale

- the placing of data into categories without any order or structure
- no numerical relationship between categories - even if numbers are used for representation
- eg. gender, nationality, language, region, etc.

2. Ordinal Scale

- the placing of data into categories such that the order of values is meaningful, but relative degree of difference is not known
- eg. ranking the features of a product on a scale of 1 to 5
- the Likert scale - psychometric scale commonly used in questionnaires

Highly Satisfied	Dissatisfied	Neutral	Satisfied	Highly Satisfied
1	2	3	4	5

3. Interval Scale

- numeric scale in which the order as well as the relative difference between values is known
- no "true zero"
- eg. temperature can be below $0°C$

4. Ratio Scale

- numeric scale with an absolute "zero"
- addition, subtraction, multiplication, and division are all valid operations
- eg. height, weight age, etc. - always measured from 0 to a maximum value

2.1.3 Measures of Central Tendency

a.k.a. Measures of Central Location

- a single value that describes a set of data by identifying the central position within that set of data

The most commonly used measures of central tendency are:

- Mean
  - arithmetic mean, commonly known as average
  - sum of all values divided by the number of values

- Median
  - arrange N data elements in order
  - if N is odd take the middle value
  - if N is even take the average of the two middle values

- Mode
  - the most frequently occuring value in a data set

The mean, median, and mode are all valid measures of central tendency, but under different conditions, some measures are more appropriate than others.

It is recommended to report trimmed mean along with mean when outliers are present. Trimmed mean excludes extreme data points from the calculation. Typically 5 of data from each end is excluded - which will give a robust estimate if the underlying distribution is symmetric.

Type of variable	Best Measure
Nominal	Mode
Ordinal	Median
Interval / Ratio (symmetric)	Mean
Interval / Ratio (skewed)	Median

- for a symmetric distribution the mean is appropriate - the mean is at the center
- for a skewed distribution the median is appropriate - the mean is generally not at the center

2.1.4 Measures of Variation

a.k.a. Measures of Dispersion

- an indication of the spread of measurements
- two datasets can have similar mean but vastly different variability

The most commonly used measures of variation are:

- Range
  - simple measure of variation
  - difference between highest and lowest values
  - crude measure as it does not take into account all values
  - same units as the original values
- Inter Quartile Range
  - the range between the upper quartile and the lower quartile
  - the quartiles are the values that divide the data into four equal parts
  - the values that divide each part are the first, second, and third quantiles (Q1, Q2, and Q3)
  - same units as the original values
- Variance and Standard Deviation
  - variance is the sum of the squared deviations from the mean divided by the number of data points
  - standard deviation is the positive square root of the variance, and has the same units as the original values

2.1.5 Coefficient of Variation (CV)

- relative measure of variation

- used to compare the variability in two data sets
- standard deviation divided by the mean, usually expressed as a percentage
- higher the value of CV, the more variability
- often referred to as the relative standard deviation

2.2 Beyond Mean and Variance

2.2.1 Skewness

Describes the shape of the data. It is the lack of symmetry.

- positively skewed
  - longer right tail
  - mass of distribution concentrated on the left
  - mode $<$ median $<$ mean
- symmetric
  - both tails are equal
  - mass of distribution distributed equally
  - mode = median = mean
- negatively skewed
  - longer left tail
  - mass of distribution concentrated on the right
  - mode $>$ median $>$ mean

- if skewness > 1 or < -1 the distribution is highly skewed
- if skewness is between 1 and 0.5 or between -0.5 and -1 the distribution is moderately skewed
- if skewness is between 0.5 and -0.5 then the distribution is approximately symmetric

2.2.2 Kurtosis

Defined as the measure of peakedness. Measured relative to Normal distribution.

- mesokurtic
  - Normal distribution
- leptokurtic
  - more acute peak than the Normal distribution
- platykurtic
  - flatter peak than the Normal distribution

- kurtosis = 3 (excess = 0) is called mesokurtic
- kurtosis < 3 (excess < 0) is called platykurtic
- kurtosis > 3 (excess > 0) is called leptokurtic

2.2.3 Moments

Constants which help us in knowing the characteristics and graphic shape of data.

2.3 Bivariate Relationships

2.3.1 Describing a Bivariate Relationships

- univariate data - data having one variable
- bivariate data - data having two variables
  - two numeric variables
  - two categorical variables
  - one numeric variable, one categorical variable

can be described using:

- scatter plot
- correlation coefficient
- simple linear regression

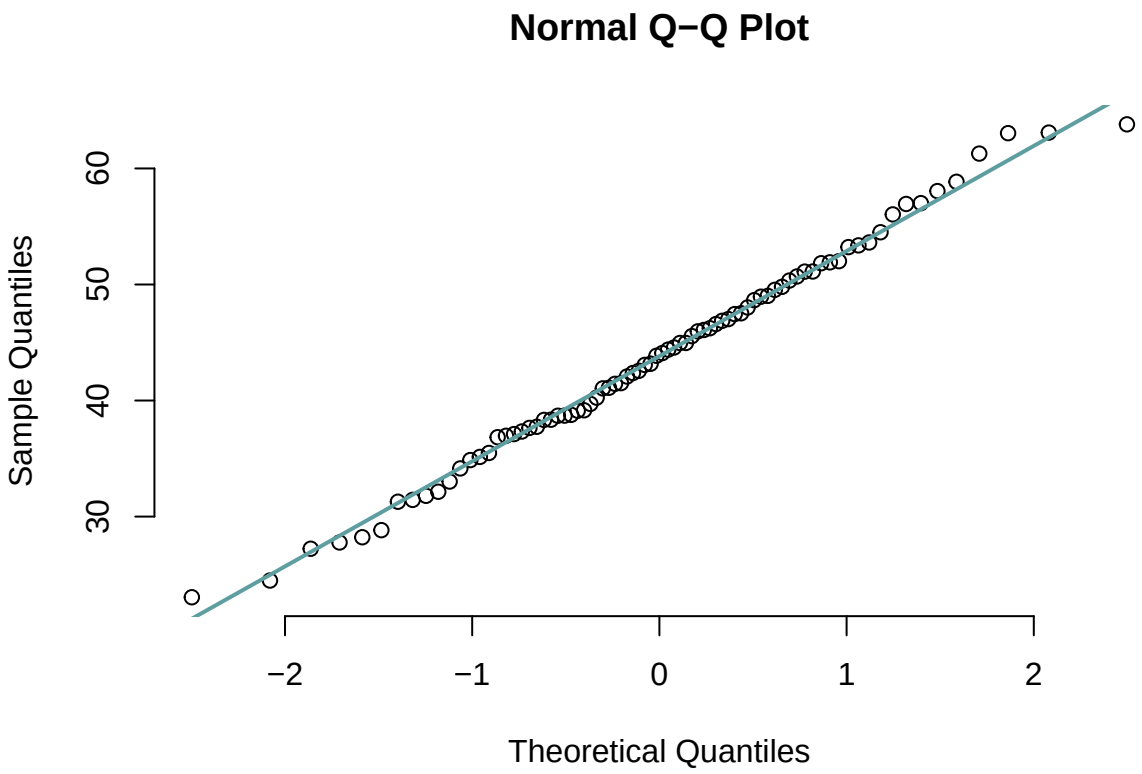
2.3.2 Scatter Plot

A scatter plot consists of:

- x-axis - the values of one variable
- y-axis - the values of another variable
- series of dots / observations
- used with two numeric continuous variables

interpreting a scatter plot:

- positive correlation
- negative correlation
- no correlation



2.3.3 Pearson's Correlation Coefficient

numerically measures the strength of a linear relation between two variables.

$$-1 \leq r \leq 1$$

Positive Correlation	$r > 0$
Negative Correlation	$r < 0$
No Correlation	$r = 0$

- the two variables can be measured in different units
- not affected by change of origin / scale

2.3.4 Line of Best Fit

- a straight line that best represents the data on a scatter plot
- may pass through some, none, or all of the points

2.3.5 Simple Linear Regression

The equation of line of best fit used to describe the relationship between two variables

mathematical form:

$$\mathbf{Y} = a\mathbf{X} + b + e$$

where:

- a - the slope
- b - the intercept
- e - the error

a and b are population parameters which are estimated using samples.

- $\mathbf{Y}$ is the dependent variable
- $\mathbf{X}$ is the independent variable

3 Data Visualisation

3.1 Chapter 1

To be completed...