1495 UNIVERSITY OF ABERDEEN | CELEBRATING 525 YEARS 1495 – 2020

ABERDEEN 2040

Revision – Week 1

Data Mining & Visualisation

Lecture 15

2025

Today...

- Exam-style questions that cover the past week's lectures

- We will walk through each one

Levels of Measurement

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

i. Country of origin
ii. Temperature in Celsius
iii. Temperature in Kelvin
iv. Exam grade: {A; B; C; D; F}
v. Age
vi. Quality of food: {bad; neither bad nor good; good}

Recall…

- Data is often grouped into one of four levels, indicating its precision

Nominal → Ordinal → Interval → Ratio

Levels of Measurement: Nominal

- Unordered classes (categorical). Data can only be categorised.
- May be coded into numeric 'dummy variables'
- Examples: Gender, race, degree program

Levels of Measurement: Ordinal

- Ordered classes (categorical). Data can be categorised and ranked.
- May also be coded into numeric variables
- Examples: Age group, educational level

Levels of Measurement: Interval

- Numerical (quantitative) data, with equal intervals but with no absolute zero
- Examples: Temperature (Celsius), year

Levels of Measurement: Ratio

- Numerical (quantitative) data, with equal intervals and with an absolute zero
- Examples: Age, weight, height, tempeature (Kelvin)

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

i. Country of origin
ii. Temperature in Celsius
iii. Temperature in Kelvin
iv. Exam grade: {A; B; C; D; F}
v. Age
vi. Quality of food: {bad; neither bad nor good; good}

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

(Nominal) i. Country of origin
ii. Temperature in Celsius
iii. Temperature in Kelvin
iv. Exam grade: {A; B; C; D; F}
v. Age
vi. Quality of food: {bad; neither bad nor good; good}

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

(Nominal) i. Country of origin

(Interval) ii. Temperature in Celsius

iii. Temperature in Kelvin

iv. Exam grade: {A; B; C; D; F}

v. Age

vi. Quality of food: {bad; neither bad nor good; good}

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

(Nominal) i. Country of origin

(Interval) ii. Temperature in Celsius

(Ratio) iii. Temperature in Kelvin

iv. Exam grade: {A; B; C; D; F}

v. Age

vi. Quality of food: {bad; neither bad nor good; good}

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

(Nominal) i. Country of origin

(Interval) ii. Temperature in Celsius

(Ratio) iii. Temperature in Kelvin

(Ordinal) iv. Exam grade: {A; B; C; D; F}

v. Age

vi. Quality of food: {bad; neither bad nor good; good}

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

(Nominal) i. Country of origin

(Interval) ii. Temperature in Celsius

(Ratio) iii. Temperature in Kelvin

(Ordinal) iv. Exam grade: {A; B; C; D; F}

(Ratio) v. Age

vi. Quality of food: {bad; neither bad nor good; good}

Levels of Measurement

Statistically, we define four levels of measurement for data: Nominal, Ordinal, Interval, and Ratio. Consider the following attributes, and point out which levels they belong to:

(Nominal) i. Country of origin

(Interval) ii. Temperature in Celsius

(Ratio) iii. Temperature in Kelvin

(Ordinal) iv. Exam grade: {A; B; C; D; F}

(Ratio) v. Age

(Ordinal) vi. Quality of food: {bad; neither bad nor good; good}

Descriptive Statistics

Descriptive Statistics

Consider the following samples:

1	3	2	1	2	3

Calculate the **mean**, **standard deviation**, and **median** for the above samples.

Then calculate the **z-score** for **each** of the above samples.

Descriptive Statistics

Recall that for samples:

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

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

If n is odd, $\mathrm{med}(x) = x_{(n+1)/2}$

If n is even, $\mathrm{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$

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

Descriptive Statistics – Mean

Consider the following samples:

1	3	2	1	2	3

Recall that for samples:

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

Descriptive Statistics – Mean

Consider the following samples:

1	3	2	1	2	3

Recall that for samples:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(1+3+2+1+2+3)}{6}$$

Descriptive Statistics – Mean

Consider the following samples:

1	3	2	1	2	3

Recall that for samples:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(1+3+2+1+2+3)}{6} = \frac{12}{6}$$

Descriptive Statistics – Mean

Consider the following samples:

1	3	2	1	2	3

Recall that for samples:

$$\bar{x} = \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i = \frac{(1+3+2+1+2+3)}{6} = \frac{12}{6} = 2$$

Descriptive Statistics – Mean

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$

Recall that for samples:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{(1+3+2+1+2+3)}{6} = \frac{12}{6} = 2$$

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$

Recall that for samples:

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

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$

Recall that for samples:

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

$$s = \sqrt{\frac{(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2+(3-2)^2}{6-1}}$$

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$

Recall that for samples:

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

$$s = \sqrt{\frac{(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2+(3-2)^2}{6-1}}$$

$$s = \sqrt{\frac{1+1+0+1+0+1}{5}}$$

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$

Recall that for samples:

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

$$s = \sqrt{\frac{(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2+(3-2)^2}{6-1}}$$

$$s = \sqrt{\frac{1+1+0+1+0+1}{5}} = \sqrt{\frac{4}{5}}$$

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$

Recall that for samples:

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

$$s = \sqrt{\frac{(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2+(3-2)^2}{6-1}}$$

$$s = \sqrt{\frac{1+1+0+1+0+1}{5}} = \sqrt{\frac{4}{5}} = \sqrt{0.8}$$

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$

Recall that for samples:

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

$$s = \sqrt{\frac{(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2+(3-2)^2}{6-1}}$$

$$s = \sqrt{\frac{1+1+0+1+0+1}{5}} = \sqrt{\frac{4}{5}} = \sqrt{0.8} = 0.89$$

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$

Recall that for samples:

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

$$s = \sqrt{\frac{(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2+(3-2)^2}{6-1}}$$

$$s = \sqrt{\frac{1+1+0+1+0+1}{5}} = \sqrt{\frac{4}{5}} = \sqrt{0.8} = 0.89$$

Descriptive Statistics – Standard Deviation

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$

Recall that for samples:

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

$$s = \sqrt{\frac{(1-2)^2+(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2+(3-2)^2}{6-1}}$$

$$s = \sqrt{\frac{1+1+0+1+0+1}{5}} = \sqrt{\frac{4}{5}} = \sqrt{0.8} = 0.89$$

Note

In the exam, you will not need a calculator.

Any numbers that you calculate will be easier to do mentally than this

Descriptive Statistics – Median

Consider the following samples:

1	3	2	1	2	3

$\bar{x} = 2$
$s = 0.89$

Recall:

If n is odd, $\mathrm{med}(x) = x_{(n+1)/2}$

If n is even, $\mathrm{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$

Descriptive Statistics – Median

(reordered)

Consider the following samples:

1	1	2	2	3	3

$\bar{x} = 2$

$s = 0.89$

Recall:

If n is odd, $\mathrm{med}(x) = x_{(n+1)/2}$

If n is even, $\mathrm{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$

Descriptive Statistics – Median

(reordered)

Consider the following samples:

1	1	2	2	3	3

$$\bar{x} = 2$$
$$s = 0.89$$

Recall:

$$\text{If } n \text{ is even, } \operatorname{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

$$\operatorname{med}(x) = \frac{x_{(6/2)} + x_{((6/2)+1)}}{2}$$

Descriptive Statistics – Median

(reordered)

Consider the following samples:

1	1	2	2	3	3

$\bar{x} = 2$

$s = 0.89$

Recall:

$$\text{If } n \text{ is even, } \operatorname{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

$$\operatorname{med}(x) = \frac{x_{(6/2)} + x_{((6/2)+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2}$$

Descriptive Statistics – Median

(reordered)

Consider the following samples:

1	1	2	2	3	3

$\bar{x} = 2$

$s = 0.89$

Recall:

$$\text{If } n \text{ is even, } \operatorname{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

$$\operatorname{med}(x) = \frac{x_{(6/2)} + x_{((6/2)+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{2+2}{2}$$

Descriptive Statistics – Median

Consider the following samples:

(reordered)

1	1	2	2	3	3

$\bar{x} = 2$

$s = 0.89$

Recall:

$$\text{If } n \text{ is even, } \mathrm{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

$$\mathrm{med}(x) = \frac{x_{(6/2)} + x_{((6/2)+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{2+2}{2} = \frac{4}{2}$$

Descriptive Statistics – Median

(reordered)

Consider the following samples:

1	1	2	2	3	3

$\bar{x} = 2$
$s = 0.89$

Recall:

$$\text{If } n \text{ is even, } \mathrm{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

$$\mathrm{med}(x) = \frac{x_{(6/2)} + x_{((6/2)+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{2+2}{2} = \frac{4}{2} = 2$$

Descriptive Statistics – Median

(reordered)

Consider the following samples:

1	1	2	2	3	3

$$\bar{x} = 2$$
$$s = 0.89$$
$$\text{med}(x) = 2$$

Recall:

$$\text{If } n \text{ is even, } \text{med}(x) = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$$

$$\text{med}(x) = \frac{x_{(6/2)} + x_{((6/2)+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{2+2}{2} = \frac{4}{2} = 2$$

Descriptive Statistics – z-score

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$
$$\mathrm{med}(x) = 2$$

Recall:

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

Descriptive Statistics – z-score

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$
$$\mathrm{med}(x) = 2$$

Recall:

$$z = \frac{(x - \bar{x})}{s} = \frac{(1 - 2)}{0.89}$$

Descriptive Statistics – z-score

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$
$$\mathrm{med}(x) = 2$$

Recall:

$$z = \frac{(x - \bar{x})}{s} = \frac{(1 - 2)}{0.89} = \frac{-1}{0.89}$$

Descriptive Statistics – z-score

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$
$$\mathrm{med}(x) = 2$$

Recall:

$$z = \frac{(x - \bar{x})}{s} = \frac{(1 - 2)}{0.89} = \frac{-1}{0.89} = -1.12$$

Descriptive Statistics – z-score

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$
$$\mathrm{med}(x) = 2$$

Recall:

$$z = \frac{(x - \bar{x})}{s} = \frac{(1 - 2)}{0.89} = \frac{-1}{0.89} = -1.12$$
$$= \frac{(2 - 2)}{0.89} = \frac{0}{0.89} = 0$$

Descriptive Statistics – z-score

Consider the following samples:

1	3	2	1	2	3

$$\bar{x} = 2$$
$$s = 0.89$$
$$\mathrm{med}(x) = 2$$

Recall:

$$z = \frac{(x - \bar{x})}{s} = \frac{(1 - 2)}{0.89} = \frac{-1}{0.89} = -1.12$$
$$= \frac{(2 - 2)}{0.89} = \frac{0}{0.89} = 0$$
$$= \frac{(3 - 2)}{0.89} = \frac{1}{0.89} = 1.12$$

Definitions and Understanding

Definitions & Understanding

- Briefly explain what you understand by the term ‘Data Mining’.
- Briefly explain what you understand by the term ‘Data Visualisation’.
- Briefly explain what you understand by the term ‘Exploratory Data Analysis’.

Definitions & Understanding

- Briefly explain what you understand by the term ‘Data Mining’.

Definitions & Understanding

- Briefly explain what you understand by the term ‘Data Mining’.

Data Mining is the process of discovering patterns and extracting useful information from data.

Definitions & Understanding

- Briefly explain what you understand by the term ‘Data Visualisation’.

Definitions & Understanding

- Briefly explain what you understand by the term ‘Data Visualisation’.

Data Visualisation is the process of designing visual representations of data. It can also be used to refer to these visual representations themselves.

Definitions & Understanding

- Briefly explain what you understand by the term ‘Exploratory Data Analysis’.

Definitions & Understanding

- Briefly explain what you understand by the term 'Exploratory Data Analysis'.
- **Exploratory Data Analysis** is the process of exploring your data. This might involve summarising or visualising aspects of data, identifying outliers, cleaning, and pre-processing.

Relationships in Data

Relationships in Data

- Briefly explain what you understand by the term ‘a weak positive correlation’.
- Briefly explain what you understand by the term ‘a strong negative correlation’.
- Briefly explain what you understand by the term ‘no correlation’.
- Briefly explain what you understand by the term ‘correlation does not imply causation’.

Relationships in Data

- Briefly explain what you understand by the term ‘a weak positive correlation’.

Correlation is a measurement of the strength of relationship between two variables.

A weak positive correlation would have a relatively small positive correlation coefficient. In other words, as variable x increases, variable y will have a slight tendency to increase.

Relationships in Data

- Briefly explain what you understand by the term ‘a strong negative correlation’.

A strong negative correlation would have a relatively large inverse correlation coefficient, representing two variables which are tightly and inversely associated with each other. In other words, as variable x increases, variable y will have a strong tendency to decrease.

Relationships in Data

- Briefly explain what you understand by the term ‘no correlation’.

No correlation would refer to a correlation coefficient close to 0. In other words, there is little to no relation between variables x and y.

Relationships in Data

- Briefly explain what you understand by the term ‘correlation does not imply causation’.

The term ‘correlation does not imply causation’ refers to the fact that correlation and causation are distinct terms. Correlation is a measure of the strength of an association between two variables, whereas causation is the process of one event causing or producing another event. From a given correlation coefficient, we cannot know whether causation exists or not.