

Data Management

MDM4U

Qinghao Hu

November 2, 2025

Contents

1	Unit 1	2
1.1	Lecture 1	2
1.1.1	The Fundamental or Multiplicative Counting Principle	2
1.1.2	Additive Counting Principle	2
1.2	Lecture 1.2	2
1.3	Like Term Permutations	2
1.4	Pascal Triangle	3
1.5	Venn Diagrams	3
1.6	Combination	3
2	Unit 2, Probability	4
2.1	Probability	4
2.1.1	Definitions of Probability	4
2.1.2	Definitions of other stupid stuffs	4
2.1.3	Formulas	5
2.1.4	Examples	5
2.2	Dependent Events	5
2.2.1	Definitions	5
2.3	Mutually Exclusive Events	5
2.3.1	Some boring Definitions	5
3	One Variable Statistics	7
3.1	Variables and Data	7
3.1.1	Definitions	7
3.2	One Variable Graphs	8
3.2.1	Some Definitions	8
3.3	Central Tendency	10
3.3.1	Definitions of Central Tendency	10
3.4	Standard Deviation	11
3.5	Quartiles	12
3.5.1	Definitions	12
3.5.2	Percentiles	13
3.6	Spread Grouped Data	14
3.6.1	For weighted data	14
3.7	Collect data	14
3.7.1	Anonymous and not be anonymous	14
3.7.2	Survey Questions	14
3.7.3	Questions should be avoided	14
3.7.4	Experimental vs Observational study	15
3.8	sampling	15
3.8.1	Some type of sampling	15
3.9	Bias	16

4	Two variable Statistics	17
4.1	Lecture 1: Graphs	17
4.1.1	Definitions	17
4.2	Linear Correlation	18
4.2.1	Why Scatter plot?	18
4.2.2	Some boring Definitions	18

Chapter 4

Two variable Statistics

4.1 Lecture 1: Graphs

4.1.1 Definitions

Relation

In a **relation**, the variable that you need to first is called the **independent variable**. Its value determine the value of the **dependent** variable. On the scatter plot, the independent variable is located on the **horizontal** axis and the dependent variable is located on the **vertical** axis. The title of the graph should be *dependent variable vs independent variable*

Scatter Plot

A **scatter plot** is used to determine if a correlation exists between two **numerical** variables. An **outlier** is a data point that does not fit the pattern of the other data.

Line of best fit

A **line of the best fit** can be used to model the data on a scatter plot whose points follow the trend of a line. A **curve** of best fit can be used to model the data on a scatter plot whose points follow the trend of a curve. The line/curve should be solid if the data is **continuous** and dashed/dotted if the data is **discrete**

Two variable graphed

Using a scatter plot has a **positive** correlation if the trend of the data points increase from left to right. The two variables graphed using a scatter plot has a **negative** correlation if the trend of the data points decrease from left to right.

Correlation

The correlation between two variables is strong if the points on the scatter plot follow a line or a curve very closely. The correlation between two variables is **moderate** if the points on the scatter plot nearly follow a line or curve. The correlation between two variables is **weak** if the points on the scatter plot are dispersed more widely, but still show a recognizable trend.

Two variables graphed on a scatter plot shows **no** correlation if the points are so scattered that no trend is discernible.

Interpolate

To **interpolate** means to estimate values lying between given data. To interpolate from a graph means to estimate coordinates of points between those that are plotted.

extrapolate

To **extrapolate** means to estimate values lying outside the given range of data. To extrapolate from a graph means to estimate coordinates of points beyond those that are plotted.

Contingency

A **contingency** table shows the frequency or percentage distribution of two categorical variables.

4.2 Linear Correlation

4.2.1 Why Scatter plot?

There are few advantages of a scatter plot:

- Correlations
- Predictions
- Positive/negative
- Strong/weak

4.2.2 Some boring Definitions

Linear Relationship

A **linear** relationship is one in which a **change** in the independent (explanatory) variable corresponds a proportional change in the dependent (response) variable.

We can use table to calculate the correlation of two variables:

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
85	76	2	2	4	4	4
90	83	7	9	49	81	63
76	68	-7	-6	49	36	42
78	70	-5	-4	25	16	20
85	75	2	1	4	1	2
84	72	1	-2	1	4	-2
$\Sigma = 498$	$\Sigma = 444$	NA	NA	$\Sigma = 132$	$\Sigma = 142$	$\Sigma = 129$

$$\bar{x} = \frac{\sum x}{n} \quad (4.1)$$

$$\bar{y} = \frac{\sum y}{n} \quad (4.2)$$

In order to calculate the correlation coefficient of x and y , we need to get the sample standard deviation for x and y .

$$s_x = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} \quad (4.3)$$

$$s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}} \quad (4.4)$$

Then, we need to calculate the covariance of x and y :

$$s_{xy} = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{n - 1} \quad (4.5)$$

Finally, the correlation coefficient is defined by this:

$$r = \frac{s_{xy}}{s_x * s_y} \quad (4.6)$$

