The Subject Matter of Statistics

A **datum** describes a single quantity (or quality) of some object or phenomenom.

In statistics, **data**, are represented by random variables.

A **random variable** is a function, the value of which is a real number determined by the probability associated with the occurrence of each of its outcomes.

* random variables are **numeric** or **categorical**

* random variables are;
  - **continuous** if its range forms an infinite set of real numbers i.e. consists of intervals
  - **discrete** if its range is countable i.e. consists of individual (discrete) values that can be listed

Statistics is concerned with techniques for;

(1) Descriptive Statistics; the collection, presentation and summarisation of data

(2) Inferential Statistics; the analysis of data for the purpose of making decisions in the face of uncertainty

Descriptive Statistics

Descriptive statistics provide simple summaries about the data. These summaries may be;

* visual; using tables, graphs and charts to display the data, or

* quantitative; using statistical measures for

- **univariate data:** measures of central tendency (mean, median, mode) and measures of spread (variance, standard deviation, range, interquartile range)
- **bivariate data:** correlation, line of best fit

Visual Summaries

Frequency Tables

Both numerical and categorical data can be organised into a frequency table, numerical data can also be organised into a cumulative frequency table.

e.g. A class of twenty students were asked how many pets they owned. The results were;

3	0	1	4	4	1	2	0	2	2
0	2	0	1	3	1	2	1	1	3

a tally is a convenient way of recording the scores

# of pets	0	1	2	3	4	Σ
Tally	𝍸	𝍸 |	𝍸	|||	||	
Frequency	4	6	5	3	2	20

Contingency Tables

A contingency table, (two-way table), is used to present categorical data in terms of frequency counts.
They are especially useful for solving conditional probability problems.

e.g.

Eye Colour	Black	Brown	Blue	Green	Grey	Σ
Female	20	30	10	15	10	85
Male	25	15	12	20	10	82
Σ	45	45	22	35	20	167

Determine whether being male and having green eyes are independent.

A and B are independent if

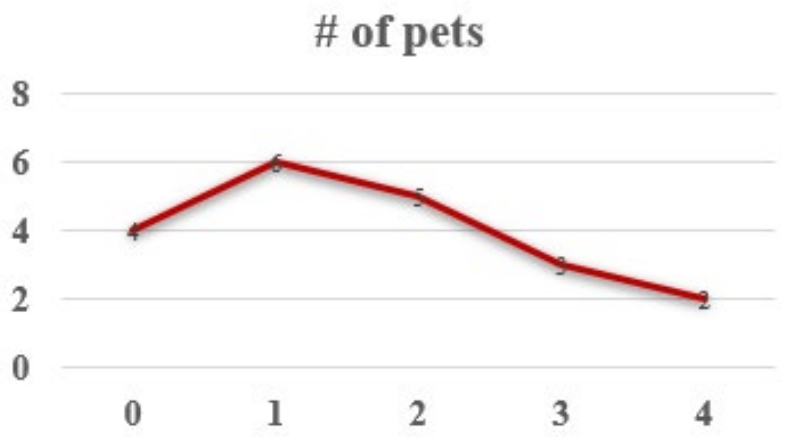
$$P(A \mid B) = P(A)$$

$$P(\text{male}) = \frac{82}{167} = 0.4910\ldots$$

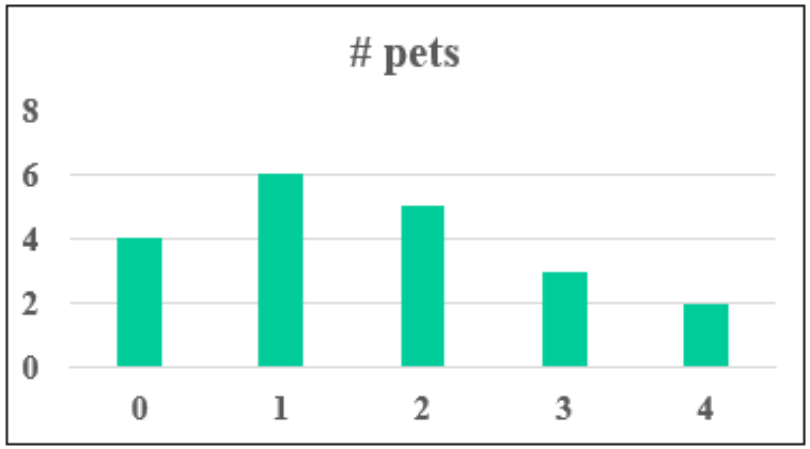
$$P(\text{male}|\text{has geen eyes}) = \frac{20}{35} = 0.5714\ldots \neq P(\text{male})$$

$\therefore$ being male and having green eyes are ***NOT*** independent

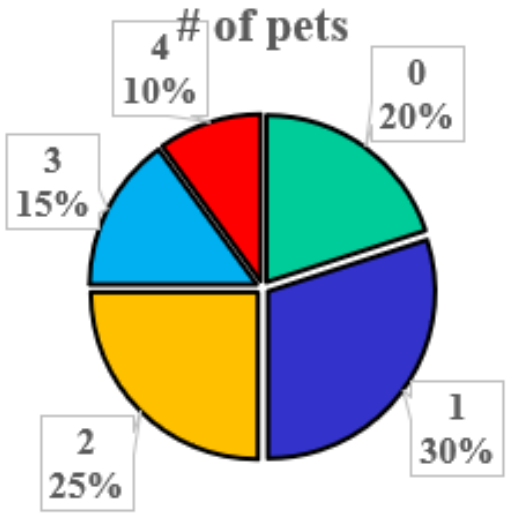
Some Common Graphs and Charts



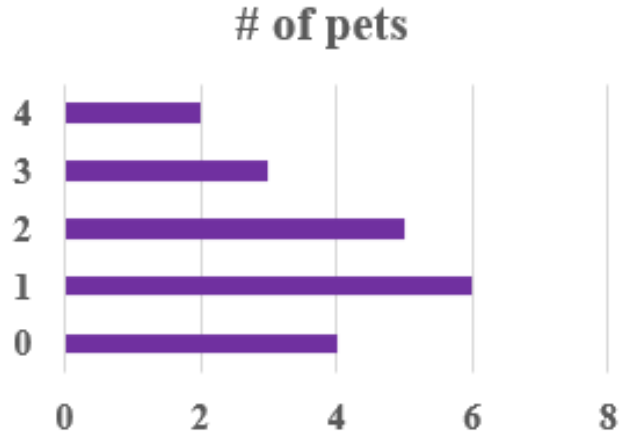
line graph



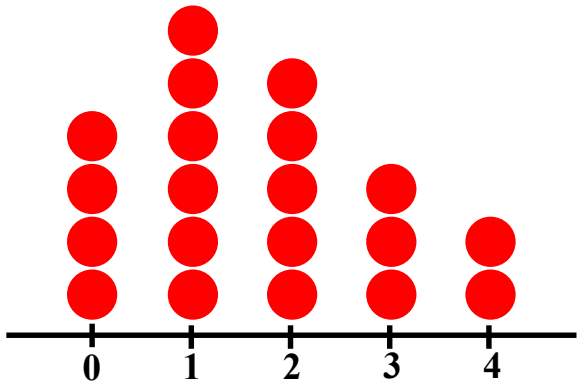
column graph



pie chart/sector graph



bar chart



dot plot

Pareto Chart

A Pareto Chart is a graph that indicates the frequency of defects, as well as their cumulative impact

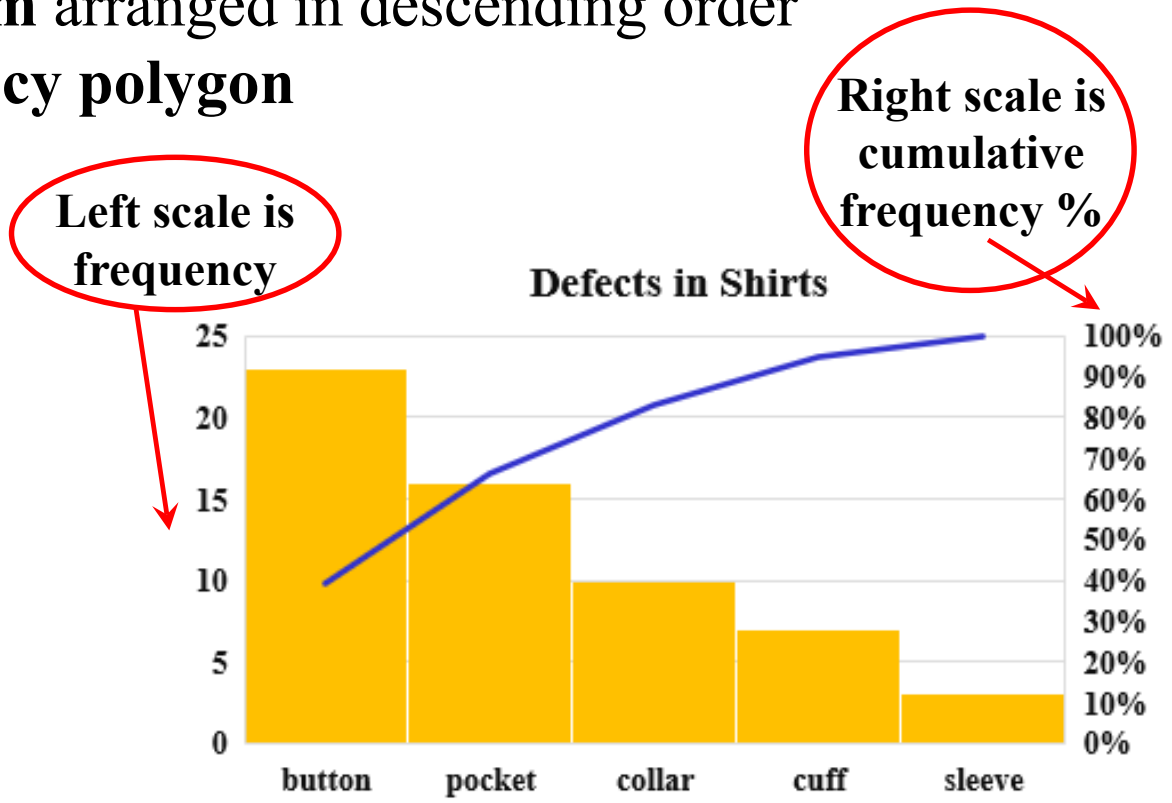
They are useful in finding the defects to prioritise in order to observe the greatest overall improvement,

It consists of two graphs drawn together;

* a **frequency histogram** arranged in descending order
* a **cumulative frequency polygon**

e.g.

Defect	frequency	Cumulative frequency
Button	23	23
Pocket	16	39
Collar	10	49
Cuff	7	56
Sleeve	3	59
Σ	59	



Eliminating the button and pocket defects will remove 66% of the defects

Quantitative Summaries

Two Measures of Central Tendency

The **mode** of a sample (dataset) is the score (datum) that occurs most often. In our example the mode is 1 pet.

The mode is not necessarily unique, the sample may be;

* bimodal (two scores with the greatest frequency)
* trimodal (three scores with the greatest frequency)
* multimodal (more than three scores with the greatest frequency)

The **median** of a sample is;

* For an odd number of scores: the middle score when the data is arranged in ascending order
* For an even number of scores: the average of the two middle scores when the data is arranged in ascending order

In our example there are 20 scores, so the median is the average of the 10^{th} and 11^{th} scores i.e. 1.5 pets.

A Measure of Spread

The **range** of a sample is the difference between the maximum and minimum score.

Only numeric data will have a range.

In our example the range is 4 – 0 = 4 pets

Exercise 15A; 1, 2ace, 3, 4, 6, 9, 10, 12ab, 14,