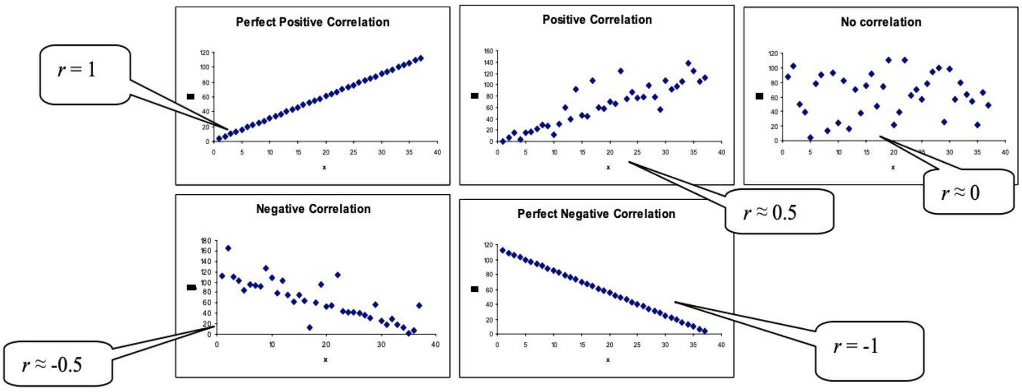
CORRELATION AND REGRESSION

Scatter Diagrams and p.m.c.c

- Product moment correlation coefficient (p.m.c.c) , **r** - number between -1 and +1 calculated to measure the correlation of a population of bivariate data
  - The closer the value of r is to +1 or -1, the stronger the correlation



$$S_{xy} = \sum(x-\bar{x})(y-\bar{y}) = \sum xy - n\bar{x}\bar{y}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}}$$

$$\text{where}: S_{xx} = \sum(x-\bar{x})^2 = \sum x^2 - n\bar{x}^2$$

$$S_{yy} = \sum(y-\bar{y})^2 = \sum_{i=1}^{n} y^2 - n\bar{y}^2$$

Rank Correlation

The Least Squares Regression Line - line of best fit which produces the least possible value of the sum of the squares of the residuals.

- Given by:

$$y-\bar{y} = \frac{S_{xy}}{S_{xx}}(x-\bar{x})$$

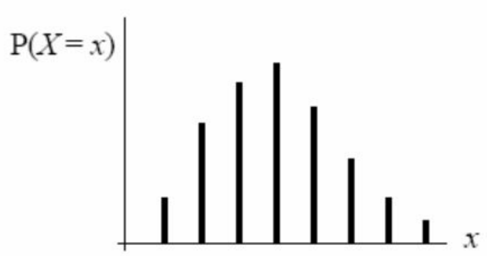
Alternatively, $y = a + bx$ where, $b = \frac{S_{xy}}{S_{xx}}, a = \bar{y} - b\bar{x}$

- (x, y), the predicted value of y is given by y^ = a + bx.
- If the regression line is a good fit to the data, the equation may be used to predict y values for x values within the given domain, i.e. interpolation.
- The corresponding residual =
- ε = y – y^ = y – (a + bx)
- The sum of the residuals = ∑ε = 0
- The least squares regression line minimises the sum of the squares of the residuals, ∑ε 2 .

DISCRETE RANDOM VARIABLES

Discrete random variables

- **Discrete random variables** with probabilities p1, p2, p3, p4, ..., pn can illustrated using a vertical line chart:



Notation:

- A discrete random variable is usually denoted by a capital letter (X, Y etc).
- Particular values of the variable are denoted by small letters (r, x etc)
- **P(X=r1)** means the probability that the discrete random variable X takes the value r1
- **∑P(X=r_k)** means the sum of the probabilities for all values of r, in other words ∑P(X=rk) = 1

Using tables

For a small set of values it is often convenient to list the probabilities for each value in a table.

r_i	r_1	r_2	r_3	....	r_{n-1}	r_n
$P(X = r_i)$	p_1	p_2	p_3	....	p_{n-1}	p_n

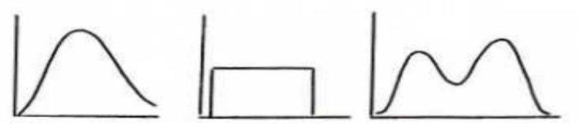
Using formulae: Sometimes it is possible to define the probability function as a formula, as a function of r, P(X = r) = f(r)

Calculating probabilities: Sometimes you need to be able to calculate the probability of some compound event, given the values from the table or function.

Explanation of probabilities: Often you need to explain how the probability P(X = rk), for some value of k, is derived from first principles.

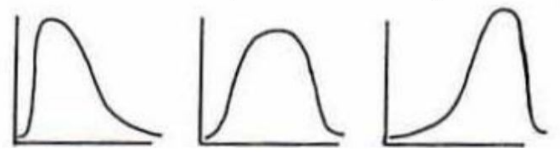
Shapes of distributions

Symmetrical (Unimodal) **Uniform** **Bimodal** *bimodal does not mean that the peaks have to be the same height*



Skew

Positive Skew **Symmetrical** **Negative Skew**



Central Tendency (averages)

Mean: $\bar{x} = \dfrac{\Sigma x}{n}$ (raw data) $\bar{x}$ **=** $\dfrac{\Sigma xf}{\Sigma f}$ (grouped data)

Median: mid-value when the data are placed in rank order

Mode: most common item or class with the highest frequency

Mid-range: (minimum + maximum) value ÷ 2

Dispersion (spread)

Range: maximum value – minimum value

Sum of squares:

$S_{xx} = \Sigma(x - \bar{x})^2 \equiv \Sigma x^2 - n\bar{x}^2$ (raw data)

$S_{xx} = \Sigma(x - \bar{x})^2 f \equiv \Sigma x^2 f - n\bar{x}^2$ (frequency dist.)

Mean square deviation: msd = $\dfrac{S_{xx}}{n}$

Root mean squared deviation: *rmsd* = $\sqrt{\dfrac{S_{xx}}{n}}$

Variance: $s^2 = \dfrac{S_{xx}}{n-1}$ **Standard deviation**: $s = \sqrt{\dfrac{S_{xx}}{n-1}}$

EXPLORING DATA

Types of data

- **Categorical data or qualitative data** are data that are listed by their properties e.g. colours of cars.
- **Numerical or quantitative data**
- **Discrete data** are data that can only take particular numerical values. e.g. shoe sizes.
- **Continuous data** are data that can take any value. It is often gathered by measuring e.g. length, temperature.

Frequency DIstributions

- **Frequency distributions:** data are presented in tables which summarise the data. This allows you to get an idea of the shape of the distribution.
- **Grouped discrete data** can be treated as if it were continuous, e.g. distribution of marks in a test.

Stem and leaf diagrams

- A concise way of displaying discrete or continuous data (measured to a given accuracy) whilst retaining the original information.
- **Visual example:**

Average daily temperatures in 16 cities are recorded in January and July. The results are

January: 2, 18, 3, 6, -3, 23, -5, 17, 14, 29, 28, -1, 2, -9, 28, 19

July: 21, 2, 16, 25, 5, 25, 19, 24, 28, -1, 8, -4, 18, 13, 14, 21

Draw a back to back stem and leaf diagram and comment on the shape of the distributions.

Answer

Jan		July
9 5 3 1	**-0**	1 4
6 3 2 2	**0**	2 5 8
9 8 7 4	**10**	3 4 6 8 9
9 8 8 3	**20**	1 1 4 5 5 8

The January data is uniform but the July data has a negative skew

Outliers

- These are pieces of data which are at least two standard deviations from the mean
  - i.e. beyond $\bar{x} \pm 2s$

Linear coding

- If the data are coded as y = ax +b then the mean and standard deviation have the coding y = a x + b (the same coding) and sy = asx (multiply by the multiplier of x)

Example:

- For two sets of data x and y it is found that they are related by the formula y = 5x – 20:

Given $\bar{x}$ = 24.8 and s_x = 7.3, find the values of $\bar{y}$ and s_y

$\bar{y}$ = (5 × 24.8) – 20 = 102

s_y = 5 × 7.3 = 36.5

PROBABILITY

The experimental probability

- The **experimental probability** of an event is = number of successes number of trials If the experiment is repeated 100 times, then the expectation (expected frequency) is equal to n × P(A).
- **The sample space** for an experiment illustrates the set of all possible outcomes. Any event is a sub-set of the sample space. Probabilities can be calculated from first principles. Example: If two fair dice are throw

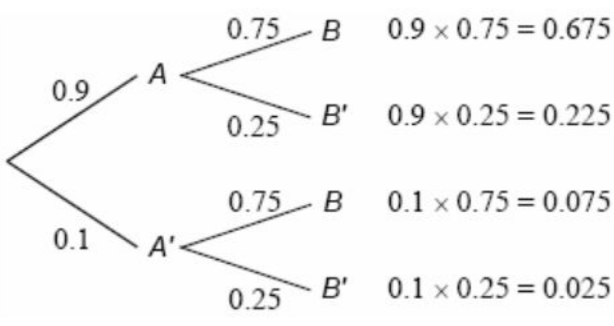
Tree Diagrams

- Multiply probabilities along the branches (AND)
- Add probabilities at the ends of branches (OR)

Example:

Event *A* (the toy is a car): P(*A*) = 0.9
Event *B* (the toy is not red): P(*B*) = 0.75

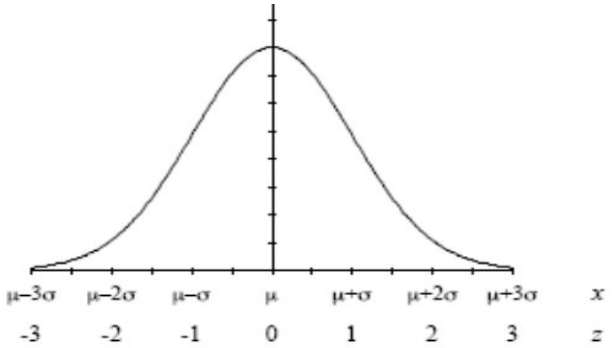
The probability of Joe getting a car that is not red is 0.675



NORMAL DISTRIBUTION

Definition

- A continuous random variable X which is bellshaped and has mean (expectation) μ and standard deviation σ is said to follow a **Normal Distribution** with **parameters** μ and σ.
- **In shorthand, X ~ N(μ, σ²)**



- **Standardised form by using transformation:**

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma z + \mu \text{, where } Z \sim \mathbf{N}(0,\ 1)$$

Calculating Probabilities

- The area to the left of the value z, representing P(Z ≤ z), is denoted by Φ(z) and is read from tables for z ≥ 0.
- Useful techniques for z ≥ 0:
  - P(Z > z) = 1 – P(Z ≤ z)
  - (Z > –z) = P(Z ≤ z)
  - P(Z < –z) = 1 – P(Z ≤ z)
- The inverse normal tables may be used to find z = Φ-1(p) for p ≥ 0.5. For p < 0.5, use symmetry properties of the Normal distribution.
  - *99.73% of values lie within 3 s.d. of the mean*

Estimating μ and/or σ

- **Use simultaneous equations of the form:**
  - x = σz + μ for matching (x, z) pairs – where z is given or may be deduced from Φ-1(p) for given value(s) of x.

Independent events

P(A and B) = P(A ∩ B) = P(A) × P(B)

Conditional probability

- If A and B are independent events then the probability that event B occurs is not affected by whether or not event A has already happened. This can be seen in example 1 above. For independent events P(B/A) = P(B)
- If A and B are dependent, as in example 2 above, then P(B/A) = $\frac{P(A \cap B)}{P(A)}$
- The multiplication law for dependent probabilities may be rearranged to give P(A and B) = P(A ∩ B) = P(A) × P(B|A)