Probability Distributions

A **random variable** is a quantity whose value depends upon chance

The **probability distribution** of a random variable is a listing of the possible values of the variable and the corresponding probabilities.

e.g. Eggs in a supermarket are sold in boxes of six. The number, X, of broken eggs in a box has the probability distribution given in the following table.

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**
$P(X = x)$	0.80	0.14	0.03	0.02	0.01	0	0

probability distribution

Properties of $P(X = x)$

$$1.\ P(X = x) \geq 0$$

$$2.\ \sum P(X = x) = 1$$

NOTE:
if $P(X=x)$ is the same for all x, then the distribution is uniform

Let X be a **discrete random variable**, then the expected value of X is;

$$\boxed{\mathrm{E}(X) = \sum xp(x)}$$

where $p(x) = \mathrm{P}(X = x) \geq 0$

Note: $\mathrm{E}(X) = \mu \quad$ (arithmetic mean)

E(X) is a measure of central tendency

e.g. If you bought 10 boxes of six, how many eggs would you expect to be broken?

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Σ**
$p(x)$	0.80	0.14	0.03	0.02	0.01	0	0	1
$xp(x)$	0	0.14	0.06	0.06	0.04	0	0	0.3

$$\begin{aligned} \mathrm{E}(X) &= \Sigma xp(x) \\ &= 0.3 \end{aligned}$$

So we would expect 0.3 eggs to be broken in each box, so in 10 boxes we would expect to have 3 broken eggs.

Let X be a **discrete random variable**, then the variance of X is;

$$\boxed{\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2}$$

where $\mathrm{E}(X^2) = \sum x^2 p(x)$

$\mu = \mathrm{E}(X)$

Note: Var(X) is a measure of spread

Standard deviation is used to measure spread using the same units as the random variable.

$$\sigma = \sqrt{\mathrm{Var}(X)}$$

or

$$\sigma^2 = \mathrm{Var}(X)$$

e.g. Find the standard deviation of the number of broken eggs in a pack of six

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Σ**
$p(x)$	0.80	0.14	0.03	0.02	0.01	0	0	1
$xp(x)$	0	0.14	0.06	0.06	0.04	0	0	0.3
$x^2p(x)$	0	0.14	0.12	0.18	0.16	0	0	0.6

$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2$

$= 0.6 - (0.3)^2$

$= 0.51$

$\sigma = \sqrt{0.51}$

$= 0.714$ (to 3 dp)

Relative Frequencies

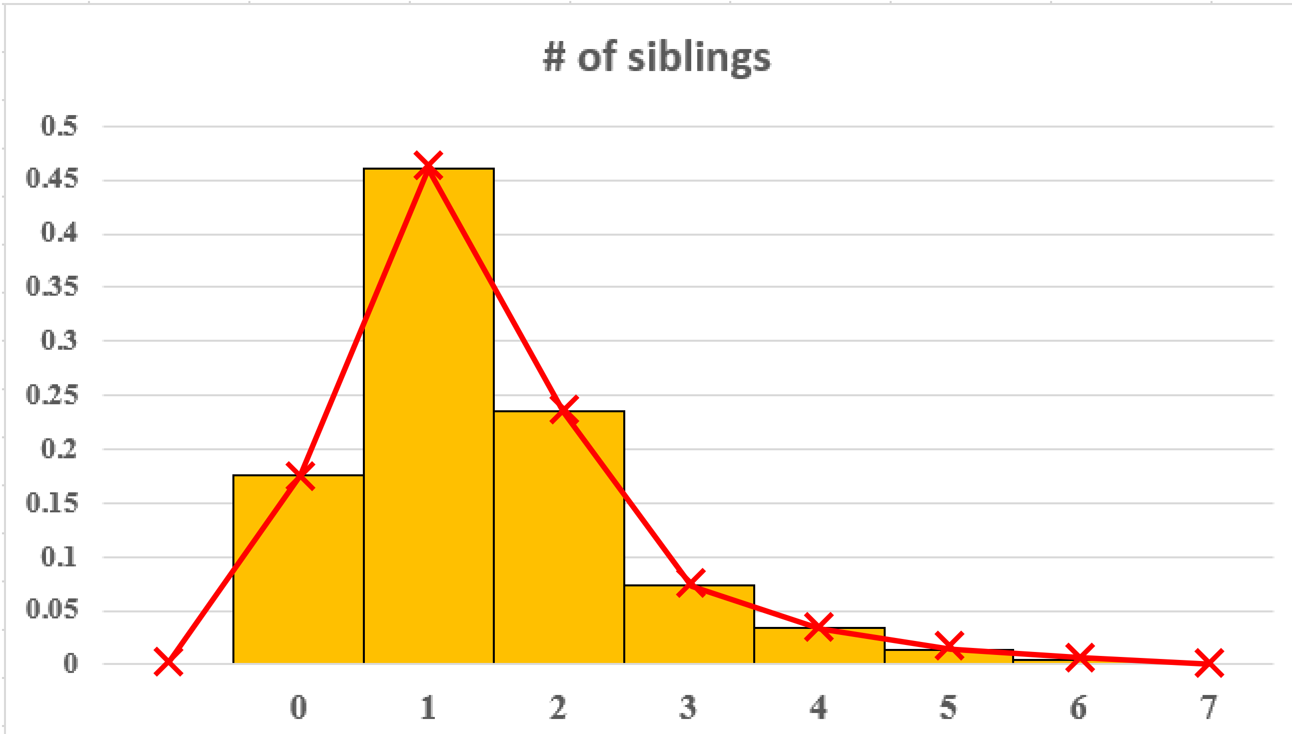
When sampling the data **relative frequencies** are used as an estimate of the theoretical probabilities.

As the size of the sample increases, these estimates will get closer to the actual theoretical probability.

e.g. Students in Year 12 were surveyed to find out the number of siblings students have. The results were placed in a relative frequency distribution table

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Σ**
f	36	94	48	15	7	3	1	204
f_r	$\frac{3}{17}$	$\frac{47}{102}$	$\frac{4}{17}$	$\frac{5}{68}$	$\frac{7}{204}$	$\frac{1}{68}$	$\frac{1}{204}$	1
cf_r	$\frac{3}{17}$	$\frac{65}{102}$	$\frac{89}{102}$	$\frac{193}{204}$	$\frac{50}{51}$	$\frac{203}{204}$	1	

a) Construct a relative frequency histogram and polygon for this data



b) Calculate the total area of the histogram rectangles

$$\text{Area} = \frac{3}{17} + \frac{47}{102} + \frac{4}{17} + \frac{5}{68} + \frac{7}{204} + \frac{1}{68} + \frac{1}{204}$$

$$= 1$$

c) Calculate the area under the relative frequency polygon

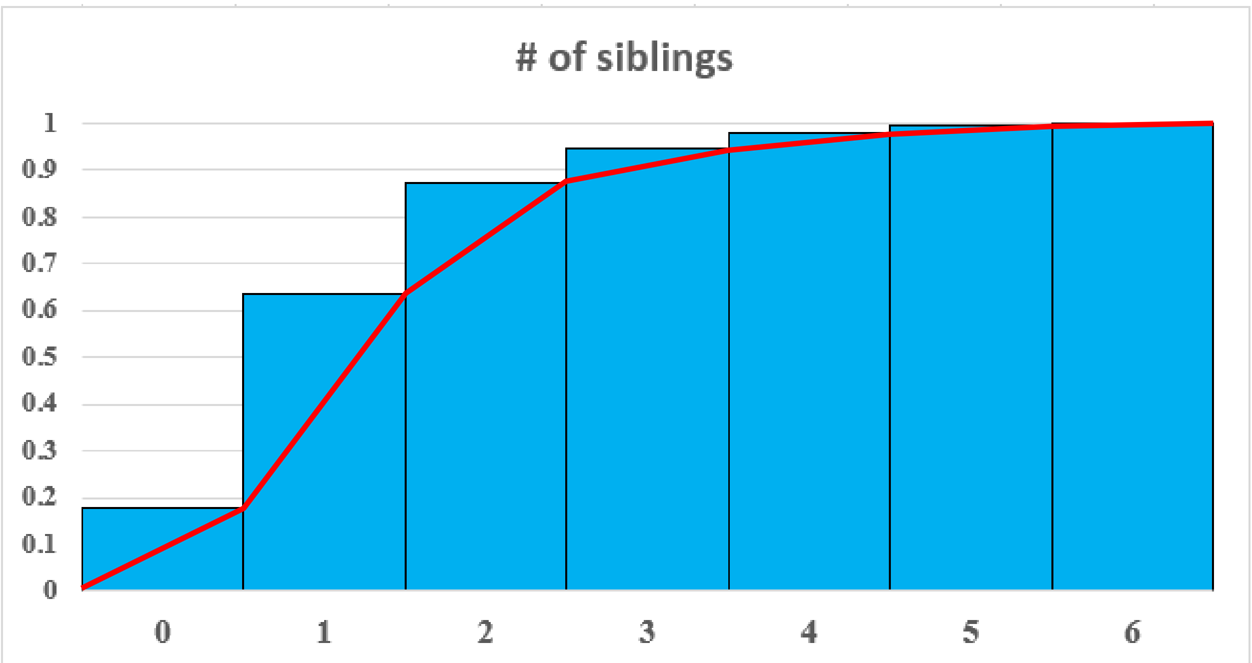
Using the trapezoidal rule

$$\text{Area} = \frac{1}{2}\left\{ 0 + 2\left(\frac{3}{17} + \frac{47}{102} + \frac{4}{17} + \frac{5}{68} + \frac{7}{204} + \frac{1}{68} + \frac{1}{204}\right) + 0 \right\}$$

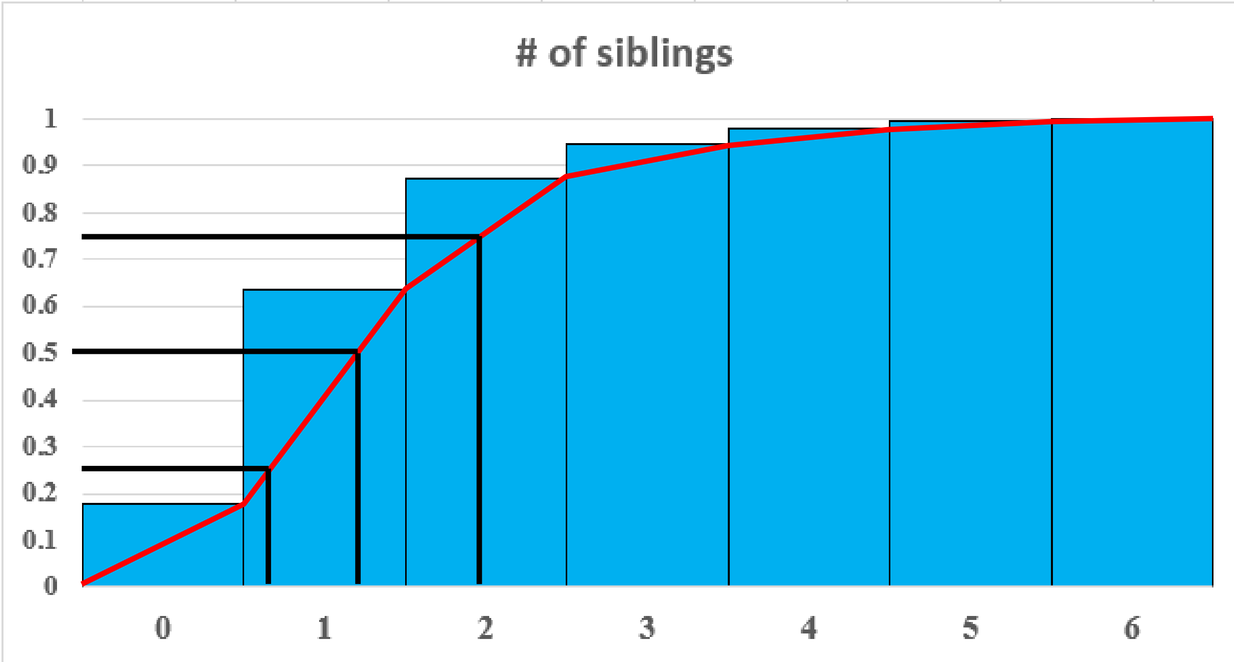
$$= 1$$

there is a direct relationship between the area under both the polygon and the area of the histogram and the total probability

d) Construct a cumulative relative frequency histogram and ogive for this data



e) Calculate the five number summary, and the interquartile range of this data.



$\min = 0$

$Q_1 = 1$

$Q_2 = 1$

$Q_3 = 2$

$\max = 6$

Five number summary: 0, 1, 1, 2, 6

IQR= 2 – 1

= 1

Exercise 16A; 1, 2, 3, 5, 7, 9, 11, 12