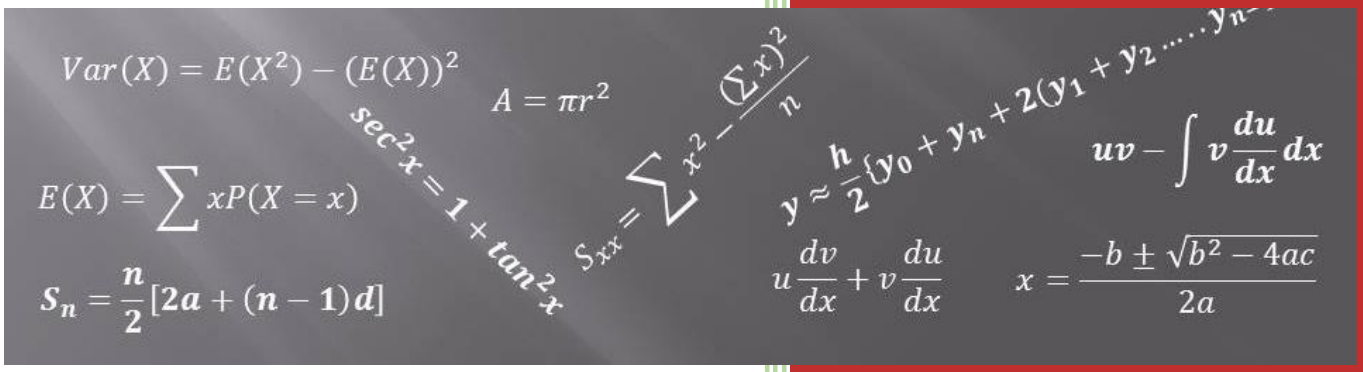
MathsGeeks

‘Everything You Need to Know’
A Level – Edexcel – S1



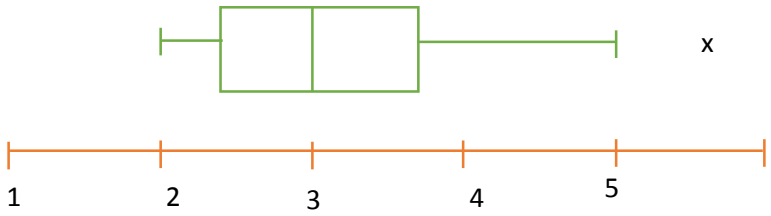
September 2013

MathsGeeks

Statistical Models are simpler, cheaper and quicker to use than a full data set for real world problems, they can be modified to new situations and allow predictions. Learn the stages:

Stage 1. The recognition of a real-world problem.
Stage 2. A statistical model is devised
Stage 3. Model used to make predictions
Stage 4. Data collected
Stage 5. Comparisons are made against the devised model.
Stage 6. Statistical concepts are used to test how well the model describes the real-world problem.
Stage 7. Model is refined

Box Plots :



The box runs from the lower quartile (2.4 above) to the upper quartile (3.7 above) with a line at the median (3.0 above). The line through the middle indicates the range of the results (2.0 to 5.0 above) and the crosses show data points which lie outside the given range and are considered to be 'outliers'.

Discrete Random Variables: remember that the total probabilities are equal to 1. Use $E(X) = \sum_{all\ x} xP(X = x)$, and $Var(X) = E(X^2) - (E(X))^2$ and learn that $E(aX + b) = aE(X) + b$ and $Var(aX + b) = a^2Var(X)$.

e.g. If random variable X has probability distribution below and $E(X) = 4.5$ set up 2 equations in p and q. They all add to 1 so $0.2 + p + 0.2 + q + 0.15 = 1$ and $p + q = 0.45$. And using $E(X) = 4.5$ then $1 \times 0.2 + 3 \times p + 5 \times 0.2 + 7 \times q + 9 \times 0.15 = 4.5$ and $3p + 7q = 1.95$. Then solve these simultaneous equations to find *p* and *q*.

x	1	3	5	7	9
P(X=x)	0.2	*p*	*0.2*	*q*	*0.15*

Skewness: Mean > median the data is positively skewed. Median < mean then the data is negatively skew. Median=Mean then no skew. $Q_3 - Q_2 > Q_2 - Q_1$ the data is positively skewed. $Q_3 - Q_2 < Q_2 - Q_1$ the data is negatively skewed. If asked which is the better approximation then the median is better where they are outliers. If the data is evenly spread then the mean is better.

Quartiles: The lower quartile is where 25% of results are below this value. The upper quartile is where 75% of results are below this value.

The Product Moment Correlation Coefficient (PMCC)

$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$ where S_{xx}, S_{yy}, S_{xy} are given in the formula book. r ranges from -1 to 1. If r is positive it represents a positive correlation and a negative number is a negative correlation. The closer the number is to 1 (or -1) the stronger the correlation. The questions require you to 'interpret' the PMCC, i.e. houses are more expensive the closer they are to the station. If the question redefines the parameters or changes the unit then r won't change as it is a ratio.

Linear Regression Line: For the line $y = a + bx$ (where a is the intercept and b is the gradient) find b using $b = \frac{S_{xy}}{S_{xx}}$ (where S_{xy} and S_{xx} are given in the formula book ($S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$ and $S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$) then find a using $a = \bar{y} - b\bar{x}$ where $\bar{y}$ and $\bar{x}$ and the mean values of y and x given by $\bar{y} = \frac{\sum y}{n}$ or $\bar{x} = \frac{\sum x}{n}$. Then substitute a and b back into $y = a + bx$.

Probabilities Learn to use $P(A \cup B) = P(A) + P(B) - (A \cap B)$. Mutually exclusive means $P(A \cap B) = 0$ so they have no intersection on a venn diagram. Events are independent if $P(A \cap B) = P(A)P(B)$. Conditional probability is the probability of A given B and is written as $P(A|B)$ and given by $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

Normal Distribution written as $X \sim N(\mu, \sigma^2)$ and $E(X) = \mu$ and $Var(X) = \sigma^2$. The tables are set up for $\emptyset(z) = P(Z < z)$ where Z is greater than μ. The equations have to be changed into this form using $P(Z < -a) = 1 - \emptyset(z)$ and $P(Z > -a) = \emptyset(a)$. Find Z using $Z = \frac{X - \mu}{\sigma}$ so e.g. for $P(X > 286)$for $X \sim N(300,25)$ then $P\left(X > \frac{286-300}{5}\right), P(Z > -2.8) = \emptyset(2.8) = 0.99744$.

Frequency Distributions and Histograms

The area of the bar on a frequency histogram is proportional to the frequency of the class and $frequency\ density = \frac{frequency}{class\ width}$. You often have to work out what area represents a frequency of one and then use this info. Learn that the mean $\mu = \frac{\sum fx}{\sum f}$ where x is the midpoint of the class and f is the frequency and that $\sigma^2 = \frac{\sum fx^2}{n} - \mu^2$and $\bar{x} = \frac{\sum fx}{n}$.

To find the median and quartiles in frequency distributions first work out which class the median or quartile falls in and then use this

$$Q_2 = lower\ class\ boundary + \frac{\frac{n}{2} - \sum f(before\ Q_2\ class)}{f\ of\ Q_2\ class} \times Q_2 class\ width$$

Replacing $\frac{n}{2}$ with $\frac{n}{4}$ for Q_1, and $\frac{3n}{4}$ for Q_3.

Don't forget to be careful with class width if it says something like to the 'nearest minute' then a class boundary of 5-9 is actually 4.5-9.5.