Candidate 1 evidence

QUESTION NUMBER	
1.(a)	Upper Fence $= Q_3 + 1.5\,IQR$ $= 13 + 1.5(13-5)$ $= 25$ Lower Fence $= Q_1 - 1.5\,IQR$ $= 5 - 1.5(13-5)$ $= -7$ As min > Lower fence & max < Upper fence there are no outliers for the 1980's data set

Candidate 2 evidence

QUESTION NUMBER	
1.(a)	$IQR = Q_3 - Q_1 = 13 - 5 = 8$ $U_{fence} = Q_3 + 1.5IQR = 13 + 1.5 \times 8 = 25$ $L_{Fence} = Q_1 - 1.5IQR = 5 - 1.5 \times 8 = -7$ So there are no outliers in the 1980s data set (ω max. = 21 < 25 & min. = 2 > -7).

Candidate 3 evidence

QUESTION NUMBER

1.(c) 1980s – data is skewed, showing that songs don't tend to stay that long in the top 40 charts, only 1 song made it past 20 weeks
– there is not much variation in the data.
– since this is a larger data set, then ~~even~~ it ~~slight~~ shows that there is not much variation

1990s – the data is ~~slightly~~ skewed
– there is some variation within the data.
– data set is ~~smaller~~ smaller.

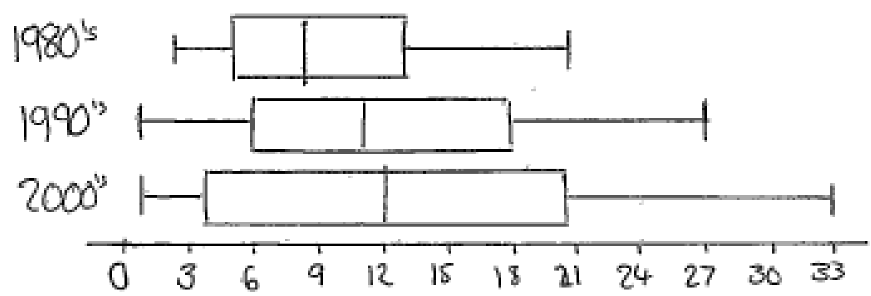
2000s – the data is quite ~~constantly~~ evely spread, meaning that there are songs that lasted long and didn't last long in the charts. There is high variation ~~however~~ in the data.

Q1	c)	as time goes on, we can see that songs stay in the ~~[illegible]~~ top charts for longer.

QUESTION NUMBER

1.(c)

Box plot to help visualise comparisons:



In general songs during the 1980's have lower values. This means that during the 1980's song stayed at the top of the charts for less time.

The 1980's also has the lowest standard deviation. This means that songs all tended to do the same amount of time in the top of the charts, unlike the 2000's with a large variance and range. This shows that there was lots of variation in the amount of time these songs stayed at the top and potentially a lot of songs that were different from the general trend.

The difference in sample sizes, shows that there were more songs in the top charts in the 1980's compared to the 1990's and 2000's. This is because there were 2% of songs picked over the same length of time, meaning the total songs is different.

Candidate 4 evidence

Candidate 5 evidence

QUESTION NUMBER

1.(c)

- there appears to be increasing variance over the years, which means songs are more likely to remain in top 40 charts for different periods of time from one another. Some may only stay for a few weeks, while others last far longer, ~~whi~~ whereas the time spent in top 40 used to be far ~~more~~ less varied.

- Some songs seem to be lasting far longer in the top 40 than they used to, which means the mean time spent in top 40 will have increased and people listen to ~~more~~ some songs more often than they used to.
- the sample size has reduced significantly during the study, which suggests that less songs make it into the top 40, but they stay there for longer periods of time.

- the data also appears to be more uniformly distributed than it used to, suggesting that some songs can be far more less successful than others.

Candidate 6 evidence

1.(c)	From 1980's to 1990's to 2000's we see an increase in the average (both mean and median) number of weeks a song stays in the top 40's, so as time progresses songs remain in the top 40 for longer on average. 1980's we see a positive skew and as the years progress the skew starts to move more negative From 1980's to 1990's to 2000's we see a decrease in sample size, showing songs remain in the top 40's for more weeks The spread of values in 1980's is much less than that of 2000's. From 1980's to 1990's to 2000's we see an increase in the spread and variability of the data showing that the number of weeks a song should expect to be in the top 40 is becoming less consistent as the years progress, with the 2000s having much more extreme values

Candidate 7 evidence

1.(c)	~~$\eta_{80's} < \mu_{80's}$~~ $\eta_{80's} < \bar{x}_{80's}$ ~~$\eta_{90's} < \mu_{90's}$~~ $\eta_{90's} < \bar{x}_{90's}$ ~~$\eta_{00's} < \mu_{00's}$~~ $\eta_{00's} < \bar{x}_{00's}$ no change all 3 decades have a positively skewed Distribution Meaning it is more likely that a song is in the charts for fewer weeks ~~$\mu_{80's} < \mu_{90's} < \mu_{00's}$~~ $\bar{x}_{80's} < \bar{x}_{90's} < \bar{x}_{00's}$ It appears there is a trend of increasing mean number of weeks - showing that songs are staying in the charts for longer as time progresses $n_{80's} > n_{90's} > n_{00's}$ There is a decrease in sample size from 80's → 00's. ~~as~~ This supports the theory that songs are staying in the charts for longer (as this means there'll be less room for other songs & therefore a smaller sample size) ~~$\sigma_{80's} < \sigma_{90's} < \sigma_{00's}$~~ $S_{80's} < S_{90's} < S_{00's}$ There is an increase in standard deviation from 80's → 00's Meaning the spread of the distribution ~~this~~ is more. This shows that the variation of length of time in the charts is increasing and so the songs are more likely to be in it for longer

Candidate 8 evidence

1.(c) The distribution of weeks spent by a chart has become gradually become more uniform.* This means that the length of time that a song will spend in the charts is becoming less predictable. *Shown by increasing IQR over the 3 decades

On average the length of time that a song will spend in the charts has increased, with the mean growing by more than three weeks and the median growing by 4 between the 80s and 00s

The decrease in Sample size ~~may~~ reflects the fact that there are less songs entering the top 40. This ~~shows that~~ reflects the fact that the number of songs staying in the charts for longer is rising as less songs are entering for the first time

Candidate 9 evidence

1.(d) (ii)	If you take a large number of samples, size n, and create a ~~100-α%~~ $(1-\alpha)100\%$ confidence interval for each one then $(1-\alpha)100\%$ of these will contain the 'population parameter' which is the true value (in this case of the mean (μ)) where α is the significance level eg. for 95% CI $\alpha = 0{\cdot}05$ $(1-0{\cdot}05)100\% = 95\%$

Candidate 10 evidence

1.(d) (ii)	The interval ~~[illegible]~~ (10.20, 15.20) will capture the true mean of the population with 95% confidence ie 95 out of 100 times

Candidate 11 evidence

1.(d)(ii)	that the true mean of this data is captured 95% of the time. i.e. the actual mean of the data will lie within the values 10.2 and 15.2 95% of the time.

Candidate 12 evidence

1.(d) (ii)	There is a 95% chance that the actual mean number of weeks a 1990s song was in the top 40 is between 10.2 and 15.2 (captured in the interval). If the data was sampled 100 times and an interval constructed, we would expect this true mean to be captured by the interval 95 times.

Candidate 13 evidence

1.(d) (ii)	~~A confidence interval is a range of values where the population mean will lie~~ A confidence interval uses a sample (smaller group from whole population) to predict where the mean (average) of that whole population is. It is a range of values in which the mean has a 95% (in the example above) chance of being.

Candidate 14 evidence

QUESTION NUMBER	
1.(e)	two tailed test p - value is high $0{\cdot}4042 > 0{\cdot}05$ $0{\cdot}4042 > 0{\cdot}10$ Accept ~~p-value~~ H_0 at 5% and 10% levels, suggesting that the true difference for number of weeks ~~to~~ in top 40 for songs in 2000 and 2010s is 0.

QUESTION NUMBER

2.(c) The teachers group is most affected by this change. This is because they have a small sample size comparatively, so just looking at numbers they don't stand out; however percentage wise it is clear they are the most likely to wear a watch.

Candidate 15 evidence

Candidate 16 evidence

2.(c)	the teachers, this is because there is a far smaller ratio of teachers to pupils, so the values need to be enhanced greatly. Less teachers in the school compared to no. of pupils. The number of teachers wearing a watch has to be multiplied to achieve percentage.

Candidate 17 evidence

2.(e)	~~Both~~ z-test for difference in population proportions $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} = \frac{79 + 54}{224 + 206} = 0{,}309$ $z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0{,}09054\ldots}{\sqrt{(0{,}309)(0{,}691)(0{,}0093)}}$ $= \underline{\underline{2{,}03}}$ ~~two tailed test~~ $P(z > 2{,}03) + P(z < -2{,}03) =$ $= 2 \times P(z > 2{,}03) = 2 \times (1 - P(z < 2{,}03))$ $= 2 \times (1 - 0{,}9788) = 0{,}0424 \approx \underline{\underline{0{,}042}}$

Candidate 18 evidence

QUESTION NUMBER

2.(e)

$H_0: p_{55} - p_{51} = 0$

$H_1: p_{55} - p_{51} \neq 0$

$p_{55} = \frac{54}{206} = 0.262 \qquad p_{51} = 0.353$

$\frac{np + np}{n+n} = \frac{54 + 79}{430}$

$p \sim N(p, \frac{pq}{n})$

$z_{crit} = 1.96$

$z = \frac{0.262 - 0.353}{\sqrt{0.309 \times 0.691\left(\frac{1}{206} + \frac{1}{224}\right)}}$

$z = \frac{-0.091}{0.0446}$

$z = \pm 2.03$

$\Phi(2.03) = 0.9788$

$\frac{1}{2}p = 1 - 0.9788$

$= 0.0212$

$p = 0.042$

Candidate 19 evidence

QUESTION NUMBER

2.(e)

Ho: Proportion of watch wearers S1 = S5

H1: proportion " " " S1 ≠ S5

Test at 5% level

$$\frac{P_1 - P_2}{\sqrt{pq\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Proportion S1 = 0·35

Proportion S5 = 0·26

$$\frac{0\cdot35 - 0\cdot23}{}$$

$$P = \frac{74 \times 0\cdot35 + 54 \times 0\cdot26}{74 + 54} = 0\cdot31$$

$$\frac{0\cdot35 - 0\cdot26}{0\cdot3\left(\frac{1}{74} + \frac{1}{54}\right)}$$

Candidate 20 evidence

QUESTION NUMBER

2.(e)

$H0: P_1 = P_2$

$H1: P_1 \cancel{\times} P_2$

$P_1 = \frac{79}{224}$. $P_2 = \frac{54}{206}$

$$\frac{P_1 - P_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$\frac{\frac{79}{224} - \frac{54}{206}}{\sqrt{0.7 \cdot 0.3\left(\frac{1}{224} + \frac{1}{206}\right)}}$$

$= 2.03$

$$P = \frac{n_1 + P_1 + n_2 P_2}{n_1 + n_2}$$

$$= \frac{224(\frac{79}{224}) + 206(\frac{54}{206})}{430}$$

$= 0.3$

$q = 1 - 0.3$

$= 0.7$

P-value

Statistics (Advanced Higher)

Candidate evidence

Question Paper 2

Candidate 21 evidence

QUESTION NUMBER

1.

H_0: No association between prevalence of infection and the sex of the fish.

H_1: There is an association.

$\frac{RT}{GT} \times CT$

O	E	$\frac{(E-O)^2}{E}$
76	104.03	7.55
129	~~100.~~ 101	7.76
399	371	2.11
322	360.03	4.02

$\sum = 21.44$

$\nu = 1$

$X^2_1 = 6.635$

$21.44 > 6.635$

So we reject H_0, meaning there is an association between prevalence of infection and sex of fish.

DO NOT WRITE IN THIS MARGIN

Candidate 22 evidence

2.(c)	
	~~Po~~ $X \sim Po(2.3)$ $Y \sim Po(1.7)$ $P(X=2) + P(Y=1.7)$ $=$ ~~0.3~~ poissonpdf $(2.3, 2)$ + poissonpdf $(1.7, 2)$ $= \underline{\underline{0.53}}$

Candidate 23 evidence

3.

$\binom{5}{2} = \frac{5!}{2!\,3!} = 10$

$P(T=0) = \frac{1}{10}$

$P(T=2) = \frac{2}{10} = \frac{1}{5}$

$P(T=4) = \frac{4}{10} = \frac{2}{5}$

$P(T=6) = \frac{2}{10} = \frac{1}{5}$

$P(T=8) = \frac{1}{10}$

t	0	2	4	6	8
P(T=t)	0,1	0,2	0,4	0,2	0,1

$E(T) = 0(0{,}1) + 2(0{,}2) + 4(0{,}4) + 6(0{,}2) + 8(0{,}1) =$

$= 4$

$V(T) = (0-4)^2\,0{,}1 + (2-4)^2\,0{,}2 + (0)^2\,0{,}4$

$+ (6-4)^2\,0{,}2 + (8-4)^2\,0{,}1 =$

$= 4{,}8$

Candidate 24 evidence

5.(b)

Let D = difference between French and German marks (French - German)

Data:

	M	M	R	E	J	K	E	S	J
French	67	83	71	59	49	89	42	55	77
German	64	82	71	62	42	85	39	50	75
Difference	3	1	0	-3	7	4	3	5	2

Assume the differences (D) are normally distributed.

$D \sim N(\mu, \sigma^2)$

$H_0: \mu = 0$

$H_1: \mu \neq 0$

Assume H_0 to be true.

$\alpha = 5\%$ 2-tailed test

we have $\bar{d} = \frac{3+1+0-3+7+4+3+5+2}{9} = 2.444$

$s_{n-1} = 2.92$

$n = 9$

Let $\bar{D}$ = mean difference in scores for 9 pupils

5.(b) cont.

$\bar{D} \sim N(\mu, \frac{\sigma^2}{9})$

$\frac{\bar{D}-\mu}{\sqrt{\frac{\sigma^2}{9}}} \sim N(0, 1^2)$

We estimate σ from s_{n-1}, and so use a t_8 distribution.

$\frac{\bar{D}-\mu}{\sqrt{\frac{s_{n-1}^2}{9}}} \sim t_8$

test statistic, $t = \frac{2.444-0}{\sqrt{\frac{2.92^2}{9}}} = 2.51141\ldots$

p-value $= 2 \times P(t_8 > 2.5114\ldots)$

$= 0.03629082\ldots$

< 0.05 $\therefore$ we have evidence to reject H_0 and conclude that there is a difference, on average, between French and German marks

Candidate 25 evidence

5.(b)

H_0: there is no difference in ~~population means~~ marks $\bar{X}_1 = \bar{X}_2$

H_1: there is a difference in ~~population means~~ marks $\bar{X}_1 \neq \bar{X}_2$

$\bar{X}_F = 65.78$

$\bar{X}_G = 63.33$

$V(x)_F = 4548.8888 - 4327.008$

$= 221.9$

$V(x)_G = 4260 - 4010.7$

$= 249.3$

$\sigma_F = 14.9$

$\sigma_G = 15.8$

$$\frac{65.78 - 63.33}{\sqrt{\frac{14.9^2}{9} + \frac{15.8^2}{9}}} = 0.3386\ldots = 0.34$$

$0.34 < 1.96$ so we don't reject H_0 meaning there is evidence that there is no difference between french and german marks.

Assume population is normally distributed.

Assume events independant of eachother.

Candidate 26 evidence

QUESTION NUMBER

7.(a)

$X \sim B(n, p)$

$E(X) = np$

$\div n \qquad \div n$

$\frac{1}{n}E(X) = \frac{np}{n}$

$E\left(\frac{X}{n}\right) = p$ as required.

$V(X) = np(1-p)$

$V(X) = npq$

$\div n^2 \qquad \div n^2$

$\frac{1}{n^2}V(X) = \frac{npq}{n^2}$

$V\left(\frac{X}{n}\right) = \frac{pq}{n}$ as required.

Candidate 27 evidence

8.(b) (i)	$\dfrac{1/5 \times 2/5}{1/5 \times 2/5 + 1/5 \times 3/5} = 0.4$

Candidate 28 evidence

QUESTION NUMBER

8.(b)(ii)

40% chance for 1 or 4

chance to lose is either $\frac{2}{5}$ or $\frac{3}{8}$

40% 37.5%

$\frac{4}{25} = 16\%$ or $\frac{6}{40} = 15\%$

15.5% chance of losing

Candidate 29 evidence

8.(b) (iii)	$P(1\|L) = \frac{0.08}{0.155} = 0.5161...$ $= \underline{0.516}$

Candidate 30 evidence

QUESTION NUMBER	
9.(a)	if $n \geq 20$, the sample mean can be approximated by the normal distribution ... if X has $E(X) = \mu$, and $V(X) = \sigma^2$, a sample mean ~~with~~ of size (n) can be approximated ... ~~the sample mean with sample size~~ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Candidate 31 evidence

QUESTION NUMBER	
9.(a)	Central limit theory states, for sufficiently large values of $n \geq 20$, then a (normal) approximation can be used $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Candidate 32 evidence

9a	The Central Limit Theorem states that the distribution of sample means from a parent population which is normally distributed is itself normally distributed. Additionally if the sample is sufficient large (the sample size $n \geq 20$) then the distribution of sample means is approximately normally distributed regardless of the distribution of the parent population. ~~The~~ for a parent population ~~...~~ with mean μ and variance σ^2 the mean of a sample of size n has mean μ and variance $\frac{\sigma^2}{n}$

Candidate 33 evidence

9.(b)

$n = 45$

$\alpha = 0.05$

≈ 1.64

$\mu = 50$

$\bar{x} = 52.6$

$\sigma^2 = 103.25$

Assume Sample means follow a normal distribution. Variables must be independent.

Ho: mean batten width = 50

H_1: mean batten width > 50

$X \sim N(50, 103.25)$

$P\left(Z > \frac{52.6-50}{\sqrt{103.25}}\right)$

$P(Z > 0.256) \sim 0.26$

$\Phi(0.26) = 0.6026$

$1.64 > 0.6026$

Evidence to reject H_1 as at the 5% significance level the mean batten with is not less than 50mm.

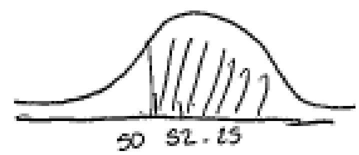
Candidate 34 evidence

9.(b)

H0 : mean battery width = 50mm $\quad x \sim N\left(52.6, \frac{103.25}{\sqrt{45}}\right)$

H1 : mean battery width > 50 mm

1tailed test at 5% $z = 1.64$



$$z > \frac{50 - 52.6}{\sqrt{\frac{103.25}{45}}} = -1.7165$$

$1 - \phi(1.17165)$

$1 - 0.8790$

$= 0.121$

1.64 > 0.121 So Accept H0, the mean battery width = 50mm.

- The wooden battens are normally distributed

Candidate 35 evidence

9.(b)	$\boxed{\begin{array}{l} H_0: \mu = 50\text{mm} \\ H_1: \mu > 50\text{mm} \end{array}}$ $\quad \alpha = 0{\cdot}05 \quad z_\alpha = 1{\cdot}64$ $n = 45 \quad \bar{x} = 52{\cdot}6 \quad \sigma^2 = 103{\cdot}25\,\text{mm}^2$ (Assume batten widths are independent) $z = \dfrac{52{\cdot}6 - 50}{103{\cdot}25} = 0{\cdot}02518\ldots$ $= 0{\cdot}025$ $0{\cdot}025 < 1{\cdot}64 \Rightarrow$ do not reject H_0 $\Rightarrow$ mean batten width is not more than 50mm

Candidate 36 evidence

10.(a)

~~$H_0: \beta = 0$~~ ~~$H_1: \beta \neq 0$~~ ~~$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$~~

~~$y = a + bx$~~ ~~$r = 0.8381$~~ ~~$r^2 = 0.7022$~~

~~$b = \frac{46.29}{278.61} = 0.1661$~~

assuming constant variance, errors are independent and mean is 0

$H_0: \beta = 0$
$H_1: \beta \neq 0$

$t = \frac{b\sqrt{S_{xx}}}{s}$ $b = 0.1661$ $s = 0.9$

$= \frac{0.1661 \times \sqrt{278.61}}{0.9} = 3.08$

CV = (5% 2 tailed df=4) = 2.776

since 3.08 > 2.776 we reject H_0 and conclude the slope parameter is not 0 so there is association between deaths and exposure at the 5% level

$b = \frac{46.29}{278.61} = 0.1661$

$s = \frac{S_{yy} - \frac{S_{xy}^2}{S_{xx}}}{n-2} = 0.9 = \frac{10.95 - \frac{46.29^2}{278.61}}{4}$

QUESTION NUMBER	
10.(a) cont.	~~$H_0: \rho = 0$~~ $H_1: \rho \neq 0$ ~~$r = 0.838$~~ ~~$r^2 = 0.702$~~ ~~$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 7.07$~~ ~~$cv = 2.776$~~ ~~$7.07 > 2.776$ so~~ ~~reject H_0~~ $H_0: \rho = 0$ $H_1: \rho = 0$ $r = 0.838$ $r^2 = 0.702$

Candidate 37 evidence

11.(a) (ii)	H_0: no difference in median reaction times H_1: difference in median reaction times. $W_m = 89$ $W = \min(W_m, m(m+n+1) - W_m)$ $= \min(89, 10(10+10+1) - 89)$ $= \min(89, 121)$ $= 89$ $E(W) = \frac{1}{2} m(m+n+1)$ $= \frac{1}{2}(121)$ $= 60.5$ $V(W) = \frac{1}{12} nm(m+n+1)$ $= \frac{1}{12} \times 10 \times 10(10+10+1)$ $= 175$ $SD(W) = 13.2288$ $P(E(W) \neq W_m) =$ $1 - \Phi\left(\frac{89 - 60.5}{13.2288}\right)$ $\Phi(2.1544)$ $= 1 - 0.9842$ $= 0.0158$ Reject H_0, $0.0158 < 0.05$ so evidence at 5% level that there is a difference in median reaction times.

Candidate 38 evidence

11.(b)

$A \sim N(2\cdot5, 0\cdot5)$ $\quad J \sim N(2\cdot0, 0\cdot3)$

$P(A < J) = P(J - A) > 0$

$J - A = D \quad D \sim N(0\cdot5, 0\cdot8)$

$P(J-A) > 0 = P(D > 0)$

$= P\left(Z > \frac{0 - 0\cdot5}{\sqrt{0\cdot8}}\right)$

$= P(Z > (-0\cdot56))$

$= 1 - (1 - \Phi(0\cdot56))$

$= 1 - 0\cdot2877$

$= \underline{0\cdot7123}$