Paper Reference(s)

6683/01

Edexcel GCE

Statistics S1

Gold Level G4

Time: 1 hour 30 minutes

Materials required for examination papers	**Items included with question**
Mathematical Formulae (Green)	Nil

Candidates may use any calculator allowed by the regulations of the Joint Council for Qualifications. Calculators must not have the facility for symbolic algebra manipulation, differentiation and integration, or have retrievable mathematical formulas stored in them.

Instructions to Candidates

Write the name of the examining body (Edexcel), your centre number, candidate number, the unit title (Statistics S1), the paper reference (6683), your surname, initials and signature.

Information for Candidates

A booklet 'Mathematical Formulae and Statistical Tables' is provided.
Full marks may be obtained for answers to ALL questions.
There are 7 questions in this question paper. The total mark for this paper is 75.

Advice to Candidates

You must ensure that your answers to parts of questions are clearly labelled.
You must show sufficient working to make your methods clear to the Examiner. Answers without working may gain no credit.

Suggested grade boundaries for this paper:

A*	A	B	C	D	E
57	50	43	36	29	22

1. A meteorologist believes that there is a relationship between the height above sea level, h m, and the air temperature, t °C. Data is collected at the same time from 9 different places on the same mountain. The data is summarised in the table below.

h	1400	1100	260	840	900	550	1230	100	770
t	3	10	20	9	10	13	5	24	16

[You may assume that $\sum h = 7150$, $\sum t = 110$, $\sum h^2 = 7171500$, $\sum t^2 = 1716$, $\sum th = 64\,980$ and $S_{tt} = 371.56$]

(*a*) Calculate S_{th} and S_{hh}. Give your answers to 3 significant figures.

(3)

(*b*) Calculate the product moment correlation coefficient for this data.

(2)

(*c*) State whether or not your value supports the use of a regression equation to predict the air temperature at different heights on this mountain. Give a reason for your answer.

(1)

(*d*) Find the equation of the regression line of t on h giving your answer in the form $t = a + bh$.

(4)

(*e*) Interpret the value of b.

(1)

(*f*) Estimate the difference in air temperature between a height of 500 m and a height of 1000 m.

(2)

May 2013

---

2. A group of office workers were questioned for a health magazine and $\frac{2}{5}$ were found to take regular exercise. When questioned about their eating habits $\frac{2}{3}$ said they always eat breakfast and, of those who always eat breakfast $\frac{9}{25}$ also took regular exercise.

Find the probability that a randomly selected member of the group

(a) always eats breakfast and takes regular exercise,

(2)

(b) does not always eat breakfast and does not take regular exercise.

(4)

(c) Determine, giving your reason, whether or not always eating breakfast and taking regular exercise are statistically independent.

(2)

January 2009

---

3. An agriculturalist is studying the yields, y kg, from tomato plants. The data from a random sample of 70 tomato plants are summarised below.

Yield (y kg)	**Frequency** (f)	**Yield midpoint** (x kg)
$0 \leq y < 5$	16	2.5
$5 \leq y < 10$	24	7.5
$10 \leq y < 15$	14	12.5
$15 \leq y < 25$	12	20
$25 \leq y < 35$	4	30

(You may use $\sum fx = 755$ and $\sum fx^2 = 12\,037.5$)

A histogram has been drawn to represent these data.

The bar representing the yield $5 \leq y < 10$ has a width of 1.5 cm and a height of 8 cm.

(a) Calculate the width and the height of the bar representing the yield $15 \leq y < 25$.

(3)

(b) Use linear interpolation to estimate the median yield of the tomato plants.

(2)

(c) Estimate the mean and the standard deviation of the yields of the tomato plants.

(4)

(d) Describe, giving a reason, the skewness of the data.

(2)

(e) Estimate the number of tomato plants in the sample that have a yield of more than 1 standard deviation above the mean.

(2)

May 2013 (R)

---

4. A researcher believes that parents with a short family name tended to give their children a long first name. A random sample of 10 children was selected and the number of letters in their family name, x, and the number of letters in their first name, y, were recorded.

The data are summarised as:

$$\sum x = 60, \quad \sum y = 61, \quad \sum y^2 = 393, \quad \sum xy = 382, \quad S_{xx} = 28$$

(*a*) Find S_{yy} and S_{xy}

(3)

(*b*) Calculate the product moment correlation coefficient, r, between x and y.

(2)

(*c*) State, giving a reason, whether or not these data support the researcher's belief.

(2)

The researcher decides to add a child with family name "Turner" to the sample.

(*d*) Using the definition $S_{xx} = \sum (x - \bar{x})^2$, state the new value of S_{xx} giving a reason for your answer.

(2)

Given that the addition of the child with family name "Turner" to the sample leads to an increase in S_{yy}

(*e*) use the definition $S_{xy} = \sum (x - \bar{x})(y - \bar{y})$ to determine whether or not the value of r will increase, decrease or stay the same. Give a reason for your answer.

(2)

May 2013 (R)

5.

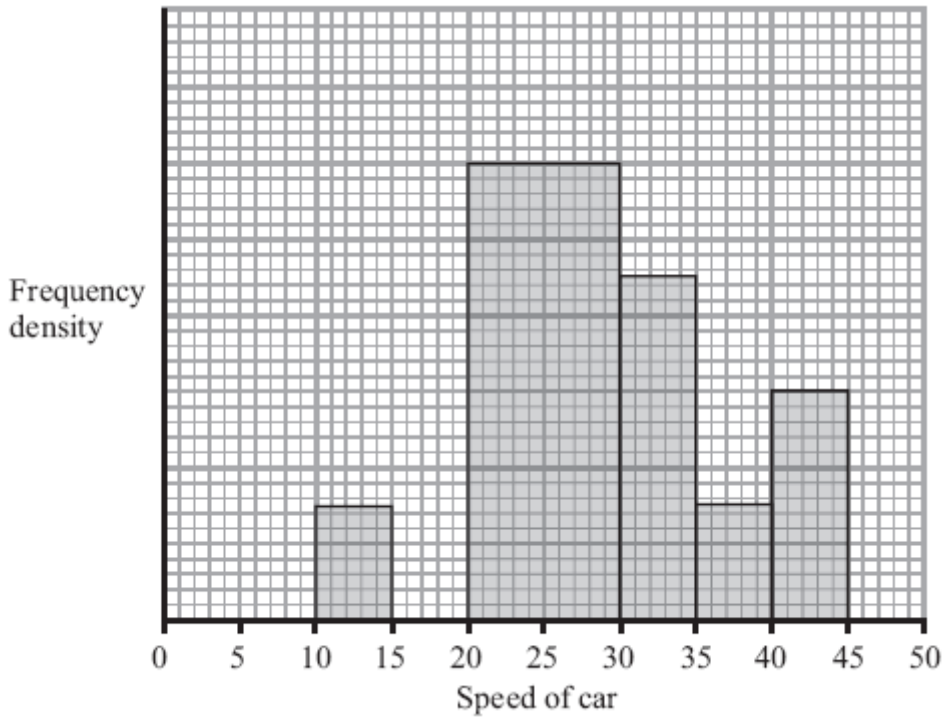

Figure 2

A policeman records the speed of the traffic on a busy road with a 30 mph speed limit.

He records the speeds of a sample of 450 cars. The histogram in Figure 2 represents the results.

(*a*) Calculate the number of cars that were exceeding the speed limit by at least 5 mph in the sample.

(4)

(*b*) Estimate the value of the mean speed of the cars in the sample.

(3)

(*c*) Estimate, to 1 decimal place, the value of the median speed of the cars in the sample.

(2)

(*d*) Comment on the shape of the distribution. Give a reason for your answer.

(2)

(*e*) State, with a reason, whether the estimate of the mean or the median is a better representation of the average speed of the traffic on the road.

(2)

May 2012

---

6. The discrete random variable X can take only the values 2, 3 or 4. For these values the cumulative distribution function is defined by

$$\mathrm{F}(x) = \frac{(x+k)^2}{25} \text{ for } x = 2, 3, 4,$$

where k is a positive integer.

(a) Find k.

(2)

(b) Find the probability distribution of X.

(3)

May 2008

---

7. The distances travelled to work, D km, by the employees at a large company are normally distributed with $D \sim \mathrm{N}(30, 8^2)$.

(a) Find the probability that a randomly selected employee has a journey to work of more than 20 km.

(3)

(b) Find the upper quartile, Q_3, of D.

(3)

(c) Write down the lower quartile, Q_1, of D.

(1)

An outlier is defined as any value of D such that $D < h$ or $D > k$ where

$$h = Q_1 - 1.5 \times (Q_3 - Q_1) \text{ and } k = Q_3 + 1.5 \times (Q_3 - Q_1).$$

(d) Find the value of h and the value of k.

(2)

An employee is selected at random.

(e) Find the probability that the distance travelled to work by this employee is an outlier.

(3)

May 2010

TOTAL FOR PAPER: 75 MARKS

END

Question Number	Scheme	Marks
1. (a)	$(S_{th}) = 64980 - \frac{7150 \times 110}{9} = -22408.9\ldots$ $\quad -\underline{\mathbf{22\ 400}}$	M1 A1
	$(S_{hh}) = 7171500 - \frac{7150^2}{9} = 1491222.2\ldots$ $\quad \underline{\mathbf{1\ 490\ 000}}$	A1
		(3)
(b)	$r = \frac{-22408.9}{\sqrt{1491222 \times 371.56}} = -0.95200068\ldots$ $\quad$ awrt $-\underline{\mathbf{0.952}}$	M1A1
		(2)
(c)	Yes as r is close to -1 (if $-1 < r < -0.5$) <u>or</u> Yes as r is close to 1 (if $1 > r > 0.5$)	B1ft
		(1)
(d)	$b = \frac{-22408.9}{1491222.2} = -0.015027...$ $\quad (\text{allow } \frac{-56}{3725})$ $\quad$ awrt -0.015	M1 A1
	$a = \frac{110}{9} - \text{"their } b\text{"} \times \frac{7150}{9} = (12.2 - -0.015 \times 794.4), = 24.1604...$	M1, A1
	so $\boldsymbol{t = 24.2 - 0.015h}$	
		(4)
(e)	0.015 is the <u>drop</u> in temp, (in ^{0}C), for every 1(m) <u>increase</u> in height above sea level.	B1
		(1)
(f)	Change = ("24.2 – 0.015" × 500) – ("24.2 – 0.015" × 1000) <u>or</u> 500 × "0.015"	M1
	$= \pm 7.5$ $\quad$ (awrt $\pm$ 7.5)	A1ft
		(2)
		[13]
2. (a)	E = take regular exercise $\quad B$ = always eat breakfast	
	$P(E \cap B) = P(E \mid B) \times P(B)$	M1
	$= \frac{9}{25} \times \frac{2}{3} = 0.24$ or $\frac{6}{25}$ or	A1
		(2)
(b)	$P(E \cup B) = \frac{2}{3} + \frac{2}{5} - \frac{6}{25}$ or $P(E' \mid B')$ or $P(B' \cap E)$ or $P(B \cap E')$	M1
	$= \frac{62}{75}$ $\quad = \frac{13}{25}$ $\quad = \frac{12}{75}$ $\quad = \frac{32}{75}$	A1
	$P(E' \cap B') = 1 - P(E \cup B) = \frac{13}{75}$ or $0.17\dot{3}$	M1 A1
		(4)
(c)	$P(E \mid B) = 0.36 \neq 0.40 = P(E)$ or $P(E \cap B) = \frac{6}{25} \neq \frac{2}{5} \times \frac{2}{3} = P(E) \times P(B)$	M1
	So E and B are <u>not</u> statistically independent	A1
		(2)
		[8]

Question Number	Scheme	Marks
3. (a)	Width = 2×1.5 = **3 (cm)**	B1
	Area = $8\times1.5=12$ cm^2 Frequency = 24 so 1 cm^2 = 2 plants (o.e.)	M1
	Frequency of 12 corresponds to area of 6 so height = **2 (cm)**	A1 (3)
(b)	$[Q_2 =]\ (5+)\ \frac{19}{24}\times5$ **or** (use of $(n+1)$) $(5+)\frac{19.5}{24}\times5$	M1
	= 8.9583… **awrt 8.96** **or** 9.0625… awrt 9.06	A1 (2)
(c)	$[\bar{x} =]\ \frac{755}{70}$ or **awrt 10.8**	B1
	$[\sigma_x =]\ \sqrt{\frac{12037.5}{70}-\bar{x}^2}$ $= \sqrt{55.6326...}$	M1A1ft
	= **awrt 7.46** (Accept s = awrt 7.51)	A1 (4)
(d)	$\bar{x} > Q_2$	B1ft
	So positive skew	dB1 (2)
(e)	$\bar{x}+\sigma \approx 18.3$ so number of plants is e.g. $\frac{(25-"18.3")}{10}\times12\ (+4)$ (o.e.)	M1
	= 12.04 so **12** plants	A1 (2)
		[13]

Question Number	Scheme	Marks
4. (a)	$S_{yy} = 393 - \frac{61^2}{10} = \underline{\mathbf{20.9}}$	M1A1
	$S_{xy} = 382 - \frac{61\times 60}{10} = \underline{\mathbf{16}}$	A1
		(3)
(b)	$[r =] \frac{"16"}{\sqrt{"20.9"\times 28}}$	M1
	$= 0.66140\ldots$ **awrt 0.661**	A1
		(2)
(c)	Researcher's belief suggests negative correlation, data suggests positive correlation	B1
	So data does not support researcher's belief	dB1
		(2)
(d)	New x equals $\bar{x} = 6$	B1
	Since $S_{xx} = \sum (x-\bar{x})^2$ the value of S_{xx} is the same = 28	dB1
		(2)
(e)	$S_{xy} = \sum (x-\bar{x})(y-\bar{y}) = \sum (x-\bar{x})y$ so the new term will be zero (since mean = x) and since S_{yy} increases	B1
	So r will decrease	dB1
		(2) **[11]**

Question Number	Scheme	Marks
5. (a)	One large square = $\frac{450}{"22.5"}$ or one small square = $\frac{450}{"562.5"}$ (o.e.)	M1
	One large square = 20 cars or one small square = 0.8 cars or 1 car = 1.25 squares	A1
	No. > 35 mph is: $4.5\times"20"$ or $112.5\times"0.8"$ (o.e. e.g. using fd)	dM1
	= **90** (cars)	A1 (4)
(b)	$[\bar{x}] = \frac{30\times12.5+240\times25+90\times32.5+30\times37.5+60\times42.5}{450}$ $\left[=\frac{12975}{450}\right]$	M1 M1
	$= 28.83...$ or $\frac{173}{6}$ **awrt 28.8**	A1 (3)
(c)	$[Q_2 =]\ 20+\frac{195}{240}\times10$ (o.e	M1
	$= 28.125$ [Use of $(n+1)$ gives 28.145…] **awrt 28.1**	A1 (2)
(d)	$Q_2 < \bar{x}$ [Condone $Q_2 \approx \bar{x}$]	B1ft
	So positive skew [so (almost) symmetric]	dB1ft (2)
(e)	[If chose skew in (d)] **median (Q_2)** [If chose symmetric in (d)] **mean ($\bar{x}$)**	B1
	Since the data is skewed or median not affected by extreme values / Since it uses all the data	dB1 (2) **[13]**
6. (a)	$F(4) = 1$, $(4+k)^2 = 25$	M1
	$k = 1$ as $k > 0$	A1 (2)
(b)		

x	2	3	4
P(X=x)	$\frac{9}{25}$	$\frac{7}{25}$	$\frac{9}{25}$

B1ft B1 B1

(3)
[5]

Question Number	Scheme	Marks
7. (a)	$P(D > 20) = P\left(Z > \frac{20-30}{8}\right)$	M1
	$= P(Z > -1.25)$	A1
	$=$ **0.8944** **awrt 0.894**	A1 (3)
(b)	$P(D < Q_3) = 0.75$ so $\frac{Q_3 - 30}{8} = 0.67$	M1 B1
	$Q_3 =$ **awrt 35.4**	A1 (3)
(c)	35.4 - 30= 5.4 so $Q_1 = 30 - 5.4 =$ **awrt 24.6**	B1ft (1)
(d)	$Q_3 - Q_1 = 10.8$ so $1.5(Q_3 - Q_1) = 16.2$ so $Q_1 - 16.2 = h$ or $Q_3 + 16.2 = k$	M1
	h=**8.4 to 8.6** and k= **51.4 to 51.6** both	A1 (2)
(e)	$2P(D > 51.6) = 2P(Z > 2.7)$	M1
	$= 2[1 - 0.9965] =$ awrt **0.007**	M1 A1 (3) **[12]**

Examiner reports

Question 1

Part (a) was, as usual, answered very well but a number of candidates lost the final mark because they did not round their answers to 3 significant figures or, more worryingly, they thought that $S_{hh} = 149$ to 3 significant figures.

Most knew how to calculate r in part (b) too but few gave a full answer to part (c). Many stated that there was negative correlation (although some thought this meant that the use of a regression equation was *not* suitable) but few stated clearly that the use of a regression equation was suitable because there was *strong* correlation. Some simply said that "the points were close to a straight line" but there was no scatter diagram to support this and without a clear statement that the strong correlation suggests this the examiners could not award the mark.

Most candidates (even those who felt that a regression equation was not appropriate!) could carry out the calculations in part (d) although a sizeable minority used S_{tt} instead of S_{hh} which gave them a somewhat unrealistic gradient of –60.3. Most found a correct gradient but often rounded their answer before calculating the intercept and the final mark was frequently lost.

Full interpretations in part (e) were rare with candidates failing to mention the drop in temperature or the rise in height above sea level or give their value. The final part was answered quite well with most candidates substituting values of 500 and 1000 into their equation, only the better candidates realized that the answer was easily found from 500b. A number of candidates seemed perfectly content with a final answer of around 30 000°C here (due to their incorrect gradient in part (d)) and lost the final mark. Candidates should be encouraged to try and engage with the context of the questions and this can help them both in interpreting their statistical calculations and assessing the reasonableness of their answers.

Question 2

This question was not answered well. It was encouraging to see many attempting to use a diagram to help them but there were often some false assumptions made and only the better candidates sailed through this question to score full marks.

The first problem was the interpretation of the probabilities given in the question. Many thought $\frac{9}{25} = \mathrm{P}(E \cap B)$ rather than $\mathrm{P}(E \mid B)$. All possible combinations of products of two of $\frac{2}{3}, \frac{2}{5}$ and $\frac{9}{25}$ were offered for part (a) but $\frac{9}{25} = \mathrm{P}(E \cap B)$ was the most common incorrect answer. In part (b) a variety of strategies were employed. Probably the most successful involved the use of a Venn diagram which, once part (a) had been answered could easily be constructed. Others tried using a tree diagram but there were invariably false assumptions made about $\mathrm{P}(E \mid B')$ with many thinking it was equal to $1 - \frac{9}{25}$. A few candidates assumed independence in parts (a) or (b) and did not trouble the scorers. The usual approach in part (c) involved comparing their answer from part (a) with the product of P(E) and P(B) although a few did use P($E|B$) and P(E). Despite the question stressing that we were looking for statistical independence here, many candidates wrote about healthy living and exercise!

The large number of candidates who confused $\mathrm{P}(E \cap B)$ and $\mathrm{P}(E \mid B)$ suggests that this is an area where students would benefit from more practice.

Question 3

In part (a) some candidates could not calculate the widths of the intervals and therefore lost all the marks. In part (b) the technique of linear interpolation is understood well but a number of candidates could not find the correct end-points. Candidates should look carefully at tables of grouped data and determine carefully the end points and widths of the intervals.

Parts (c) and (d) were answered well without quite so many false attempts at standard deviation as is often the case on S1. Part (e) was not answered so well as many candidates didn't appreciate the need to interpolate. Those who did usually arrived at the correct answer quite efficiently.

Question 4

Parts (a) and (b) were answered very well with only minor slips causing a loss of marks in a few cases. In part (c) most candidates realized there was positive correlation but some went on to state that this suggested support for the researcher's belief and only the more astute explaining that the researcher should have been expecting a negative correlation and these data therefore did not offer support. Parts (d) and (e) were challenging. In part (d) many stated that S_{xx} would remain the same but they were unable to provide an adequate reason. In part (e) most thought that r would stay the same giving the text book reason that "it is not affected by coding" but a few did realize that S_{xy} would stay the same and so the increase in S_{yy} meant that r would in fact decrease.

Question 5

This question was not answered very well. Many candidates either made a poor attempt at part (a) and then abandoned the question or just left it blank and moved on. Those who correctly formed a frequency table often scored well, whilst others who failed to complete part (a) struggled to make any headway with the remainder of the question.

In part (a) there needed to be some attempt to count squares and 22.5, 562.5 or 112.5 (small squares greater than or equal to 35 mph) were frequently seen. However many candidates did not appreciate the key idea that area is proportional to frequency and there was no attempt to combine this figure with the total frequency of 450. Those who did combine their figures and were able to come up with a correct relationship between area and number of cars (e.g. 1 large square represents 20 cars) were usually able to complete this part successfully although a few found the number of cars speeding above 30 mph instead of 35 mph as required. A few candidates stumbled upon 90 by dividing the 450 cars by the 5 bars of the histogram but this, of course, received no credit.

In part (b) most attempts tried to use mid-points but many struggled to find suitable frequencies and a few were unsure of the class widths (using 6 and 11). Some used the number of squares as frequencies but they rarely had a compatible denominator for their expression for the mean.

Most attempts at part (c) realised that interpolation was required but many promising solutions used 19.5 or 20.5 as class boundaries rather than 20.

Although they may not have had the correct values for the mean and median many had some values which they could use to answer part (d). A simple comparison of their values (e.g. mean greater than median) earned them the first mark and then an appropriate comment about the skewness (such as positive skew) the second. Some attempted to calculate the quartiles and invariably these were incorrect. Candidates should consider the amount of work involved in finding these values and compare it with the tariff for the question: 2 marks for a comment

and a reason should not involve half a page of calculations. Other candidates tried to justify their comment from the shape of the histogram ignoring the calculations in parts (b) and (c).

In part (e) we required a choice of mean or median that was compatible with their conclusion in part (d). Some candidates who had correctly concluded that the distribution was skewed in part (d) still chose the mean, because it uses all the data, but there were many correct answers seen to this part.

Question 6

This question was an excellent example of why students should revise the syllabus and not just from past papers. Only a minority of candidates tackled this question effectively; some candidates seemed to have no idea at all as to how to tackle the question. Those who gave correct solutions often made many incorrect attempts in their working. The vast majority showed an understanding of discrete random variables but most missed or did not understand the word "cumulative" and consequently spent a lot of time manipulating quadratic expressions trying to make them into a probability distribution. The majority view was that F(1) + F(2) + F(3) = 1 which led to a lot of incorrect calculations.

Question 7

This question proved to be quite challenging for a high proportion of candidates. A significant number either made no attempt at the question or offered very little in the way of creditable solutions, with many unable to progress beyond part (a). Time issues may have been a contributing factor in some cases.

The majority of candidates however, were able to earn some credit at least in part (a), for their standardisation, although whilst this was often completely correct, a fairly common mistake was to give 1 – 0.8944 = 0.1056 as their final answer.

Many students did not recognise that they needed to actually use the normal distribution in part (b) and part (c), giving rise to extremely poor attempts by numerous candidates. Of these, many merely gave 45 and 15 as their quartiles, whilst others calculated $\frac{3}{4}$ of some value as their upper quartile (for example $\frac{3}{4} \times 60$) and $\frac{1}{4}$ of the same value as their lower quartile. Alternatively, of those who understood that they were required to use the normal distribution, most attempts were successful, though there were some instances of their setting their standardisation equal to a probability, usually 0.75 or P($Z < 0.75$), and not a z-value. Unfortunately 0.68 was used fairly frequently as the z value. The majority of candidates were however able to follow through their value of the upper quartile to find their lower quartile using symmetry, though some performed a second calculation involving standardisation. Some candidates miscalculated their lower quartile as $\frac{1}{3}$ of their upper quartile.

Despite previous errors most candidates tended to be successful in substituting their values correctly into at least one of the given formulae. However, a few seemed unaware of the order of the operations.

The final part of the question also proved difficult for many candidates with some running into trouble as a consequence of previous errors in part (b), part (c) and part (d) and others providing no attempt at all. Indeed, for numerous candidates, incorrect values for h and k led to probabilities of 0 being calculated from results such as P($Z > 7$) and thus many creditable attempts lost marks through earlier inaccuracies.

Statistics for S1 Practice Paper Gold Level G4

Qu	Max Score	Modal score	Mean %	ALL	A*	A	B	C	D	E	U
1	13	11	65	8.46	11.72	11.08	9.61	8.61	7.61	6.70	4.73
2	8		33	2.62		3.82	1.98	1.53	1.31	1.01	0.64
3	13		67	8.73	10.84	10.13	8.54	7.53	6.43	4.72	3.11
4	11		59	6.50	7.82	7.18	6.29	5.70	5.56	5.14	4.50
5	13		39	5.02	10.63	9.16	5.68	3.90	2.78	2.12	1.48
6	5		25	1.23		2.87	0.98	0.55	0.32	0.16	0.06
7	12		43	5.20	10.39	8.98	6.01	4.31	3.10	2.08	0.98
	75		**50**	**37.76**		**53.22**	**39.09**	**32.13**	**27.11**	**21.93**	**15.50**