Paper Reference(s)

6683/01

Edexcel GCE

Statistics S1

Silver Level S1

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
68	61	55	49	43	38

This publication may only be reproduced in accordance with Edexcel Limited copyright policy.
©2007–2013 Edexcel Limited.

1. A personnel manager wants to find out if a test carried out during an employee's interview and a skills assessment at the end of basic training is a guide to performance after working for the company for one year.

The table below shows the results of the interview test of 10 employees and their performance after one year.

Employee	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Interview test, x %	65	71	79	77	85	78	85	90	81	62
Performance after one year, y %	65	74	82	64	87	78	61	65	79	69

[You may use $\sum x^2 = 60\,475$, $\sum y^2 = 53\,122$, $\sum xy = 56\,076$]

(*a*) Showing your working clearly, calculate the product moment correlation coefficient between the interview test and the performance after one year.

(5)

The product moment correlation coefficient between the skills assessment and the performance after one year is –0.156 to 3 significant figures.

(*b*) Use your answer to part (*a*) to comment on whether or not the interview test and skills assessment are a guide to the performance after one year. Give clear reasons for your answers.

(2)

January 2008

2. The marks of a group of female students in a statistics test are summarised in Figure 1.

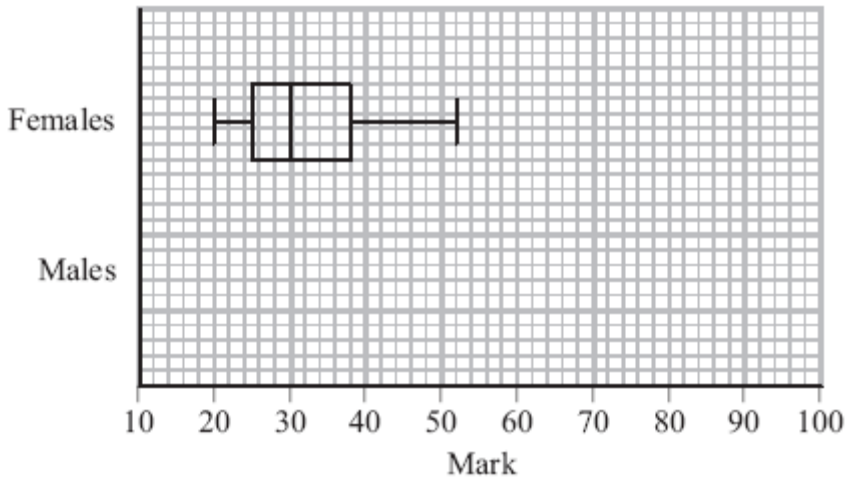


Figure 1

(*a*) Write down the mark which is exceeded by 75% of the female students.

(1)

The marks of a group of male students in the same statistics test are summarised by the stem and leaf diagram below.

Mark	(2|6 means 26)	Totals
1	4	(1)
2	6	(1)
3	4 4 7	(3)
4	0 6 6 7 7 8	(6)
5	0 0 1 1 1 3 6 7 7	(9)
6	2 2 3 3 3 8	(6)
7	0 0 8	(3)
8	5	(1)
9	0	(1)

(*b*) Find the median and interquartile range of the marks of the male students.

(3)

An outlier is a mark that is

either more than 1.5 × interquartile range above the upper quartile

or more than 1.5 × interquartile range below the lower quartile.

(*c*) On graph paper draw a box plot to represent the marks of the male students, indicating clearly any outliers.

(5)

(*d*) Compare and contrast the marks of the male and the female students.

(2)

May 2013

---

3. On a randomly chosen day the probability that Bill travels to school by car, by bicycle or on foot is $\frac{1}{2}$, $\frac{1}{6}$ and $\frac{1}{3}$ respectively. The probability of being late when using these methods of travel is $\frac{1}{5}$, $\frac{2}{5}$ and $\frac{1}{10}$ respectively.

(a) Draw a tree diagram to represent this information.

(3)

(*b*) Find the probability that on a randomly chosen day

(i) Bill travels by foot and is late,

(ii) Bill is not late.

(4)

(*c*) Given that Bill is late, find the probability that he did not travel on foot.

(4)

May 2009

---

4. The marks, x, of 45 students randomly selected from those students who sat a mathematics examination are shown in the stem and leaf diagram below.

Mark										Totals
3	6	9	9							(3)
4	0	1	2	2	3	4				(6)
4	5	6	6	6	8					(5)
5	0	2	3	3	4	4				(6)
5	5	5	6	7	7	9				(6)
6	0	0	0	0	1	3	4	4	4	(9)
6	5	5	6	7	8	9				(6)
7	1	2	3	3						(4)

Key	(3 | 6 means 36)

(*a*) Write down the modal mark of these students.

(1)

(*b*) Find the values of the lower quartile, the median and the upper quartile.

(3)

For these students $\sum x = 2497$ and $\sum x^2 = 143\ 369$.

(*c*) Find the mean and the standard deviation of the marks of these students.

(3)

(*d*) Describe the skewness of the marks of these students, giving a reason for your answer.

(2)

The mean and standard deviation of the marks of all the students who sat the examination were 55 and 10 respectively. The examiners decided that the total mark of each student should be scaled by subtracting 5 marks and then reducing the mark by a further 10 %.

(*e*) Find the mean and standard deviation of the scaled marks of all the students.

(4)

January 2012

5. On a randomly chosen day, each of the 32 students in a class recorded the time, t minutes to the nearest minute, they spent on their homework. The data for the class is summarised in the following table.

Time, t	Number of students
10 – 19	2
20 – 29	4
30 – 39	8
40 – 49	11
50 – 69	5
70 – 79	2

(*a*) Use interpolation to estimate the value of the median.

(2)

Given that

$$\sum t = 1414 \qquad \text{and} \qquad \sum t^2 = 69\,378,$$

(*b*) find the mean and the standard deviation of the times spent by the students on their homework.

(3)

(*c*) Comment on the skewness of the distribution of the times spent by the students on their homework. Give a reason for your answer.

(2)

January 2011

---

6. A travel agent sells flights to different destinations from *Beerow* airport. The distance d, measured in 100 km, of the destination from the airport and the fare £f are recorded for a random sample of 6 destinations.

Destination	*A*	*B*	*C*	*D*	*E*	*F*
d	2.2	4.0	6.0	2.5	8.0	5.0
f	18	20	25	23	32	28

[You may use $\sum d^2 = 152.09$ $\sum f^2 = 3686$ $\sum fd = 723.1$]

(*a*) On graph paper, draw a scatter diagram to illustrate this information.

(2)

(*b*) Explain why a linear regression model may be appropriate to describe the relationship between f and d.

(1)

(*c*) Calculate S_{dd} and S_{fd}.

(4)

(*d*) Calculate the equation of the regression line of f on d giving your answer in the form $f = a + bd$.

(4)

(*e*) Give an interpretation of the value of b.

(1)

Jane is planning her holiday and wishes to fly from *Beerow* airport to a destination t km away. A rival travel agent charges 5p per km.

(*f*) Find the range of values of t for which the first travel agent is cheaper than the rival.

(2)

May 2010

7. Given that

$$P(A) = 0.35 \text{ , } P(B) = 0.45 \text{ and } P(A \cap B) = 0.13,$$

find

(*a*) $P(A \cup B)$,

(2)

(*b*) $P(A' \mid B')$.

(2)

The event C has $P(C) = 0.20$.

The events A and C are mutually exclusive and the events B and C are independent.

(*c*) Find $P(B \cap C)$.

(2)

(*d*) Draw a Venn diagram to illustrate the events A, B and C and the probabilities for each region.

(4)

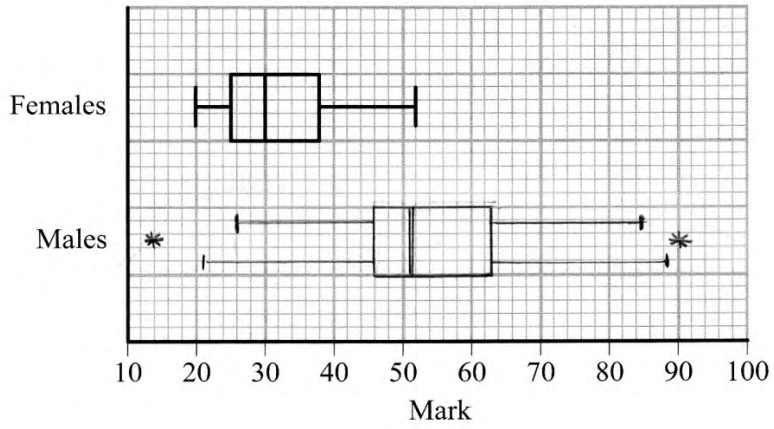
(*e*) Find $P([B \cup C]')$.

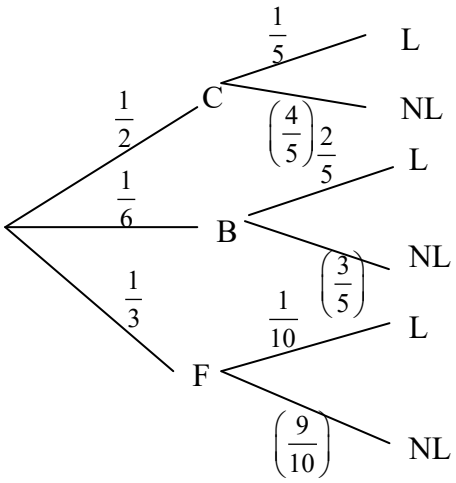
(2)

January 2013

TOTAL FOR PAPER: 75 MARKS

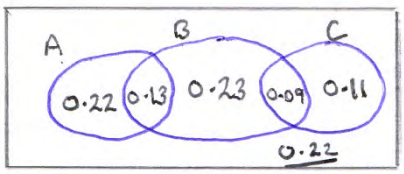
END

Question Number	Scheme	Marks
1. (a)	$\Sigma x = 773$, $\Sigma y = 724$	B1, B1
	$r = \dfrac{10\times 56076 - 773\times 724}{\sqrt{\left(10\times 60475 - 773^2\right)\left(10\times 53122 - 724^2\right)}}$ o.e.	M1 A1ft
	$r = 0.155357.....$	A1
		(5)
(b)	Both weak correlation	B1g B1h
	Neither score is a good indication of future performance	
	Interview test is slightly better since correlation is positive	
		(2)
		[7]
2. (a)	25 (allow any x where $24 < x < 26$)	B1
		(1)
(b)	$\mathbf{Q_2}$ (or median or m) = **51**	B1
	IQR = 63 – 46 ,= **17** (or $Q_3 - Q_1 = 17$)	M1, A1
		(3)
(c)	Outliers given by 46 – 1.5 ×17 = 20.5 or 63 + 1.5 ×17 = 88.5	M1
	Outliers limits are **20.5** and **88.5**	A1
		
	Allow lower whisker to 20.5 and upper whisker to 88.5	M1
	Do **not** allow a mix of whiskers e.g 20.5 and 85	A1ft
	Do **not** allow both sets of whiskers	B1
		(5)
(d)	**Medians:** Median for females lower than males	
	IQR: IQR for females smaller than males. Allow "lower/higher" but not "wider"	B1ft
	Range: Range of females is less than males	B1ft
	Skewness: Male and female marks are both positively skew	
		(2)
		[11]

Question Number	Scheme	Marks
3. (a)	 Correct tree All labels Probabilities on correct branches	B1 B1 B1 (3)
(b)(i)	$\frac{1}{3}\times\frac{1}{10}=\frac{1}{30}$ or equivalent	M1 A1 (2)
(ii)	$\frac{1}{3}\times\frac{1}{10}=\frac{1}{30}$ or equivalent	M1
	$\text{CNL} + \text{BNL} + \text{FNL} = \frac{1}{2}\times\frac{4}{5}+\frac{1}{6}\times\frac{3}{5}+\frac{1}{3}\times\frac{9}{10}$	A1 (2)
(c)	$\text{P}(F'/L)=\frac{\text{P}(F'\cap L)}{\text{P}(L)}$ Attempt correct conditional probability	M1
	$=\frac{\frac{1}{6}\times\frac{2}{5}+\frac{1}{2}\times\frac{1}{5}}{1-(ii)}$ $\frac{\text{numerator}}{\text{denominator}}$	$\frac{\text{A1}}{\text{A1ft}}$
	$=\frac{\frac{5}{30}}{\frac{1}{5}}=\frac{5}{6}$ or equivalent cao	A1 (4) **[11]**

Question Number	Scheme	Marks
4. (a)	60	B1 (1)
(b)	$Q_1 = 46$	B1
	$Q_2 = 56$	B1
	$Q_3 = 64$	B1 (3)
(c)	mean = 55.48…. or $\frac{2497}{45}$ awrt 55.5	B1
	sd = $\sqrt{\frac{143369}{45} - \left(\frac{2497}{45}\right)^2}$	M1
	= 10.342… (s = 10.459..)	A1 (3)
(d)	Mean < median < mode or $Q_2 - Q_1 > Q_3 - Q_2$ with or without their numbers or median closer to upper quartile (than lower quartile) or (mean-median)/sd <0;	B1
	negative skew;	B1dep (2)
(e)	mean = $(55-5)\times 0.9$	M1
	= 45	A1
	sd = 10×0.9	M1
	= 9	A1 (4) **[13]**
5. (a)	Median = 32/2 = 16th term (16.5)	
	$\frac{x-39.5}{49.5-39.5} = \frac{16-14}{25-14}$ or $x = 39.5 + \left(\frac{2}{11}\times 10\right)$	M1
	Median = 41.3 (use of n + 1 gives 41.8) **(awrt 41.3)**	A1 (2)
(b)	Mean = $\frac{1414}{32}$ = 44.1875 **(awrt 44.2)**	B1
	Standard deviation = $\sqrt{\frac{69378}{32} - \left(\frac{1414}{32}\right)^2}$	M1
	= 14.7 (or s = 14.9)	A1 (3)
(c)	mean > median therefore positive skew	B1ft B1ft (2) **[7]**

Question Number	Scheme	Marks
6. (a)	See overlay	B1 B1 (2)
(b)	The **points** lie reasonably close to a straight **line** (o.e.)	B1 (1)
(c)	$\sum d = 27.7, \quad \sum f = 146$ (both, may be implied)	B1
	$S_{dd} = 152.09 - \frac{(27.7)^2}{6} = 24.208\ldots..$ **awrt <u>24.2</u>**	M1 A1
	$S_{fd} = 723.1 - \frac{27.7 \times 146}{6} = 49.06\ldots.$ **awrt <u>49.1</u>**	A1 (4)
(d)	$b = \frac{S_{fd}}{S_{dd}} = 2.026\ldots.$ **awrt <u>2.03</u>**	M1 A1
	$a = \frac{146}{6} - b \times \frac{27.7}{6} = 14.97\ldots..$ so ***<u>f = 15.0 + 2.03d</u>***	M1 A1 (4)
(e)	A flight costs **£2.03 (or about £2)** for every extra **100km** or about **2p** per **km**.	B1ft (1)
(f)	$15.0 + 2.03d < 5d$ so $d > \frac{15.0}{(5 - 2.03)} = 5.00 \sim 5.05$	M1
	So $\boldsymbol{t} >$ 500~505	A1 (2) **[14]**

Question Number	Scheme	Marks
7. (a)	$P(A \cup B) = 0.35 + 0.45 - 0.13$ or $0.22 + 0.13 + 0.32$ $= \mathbf{0.67}$	M1 A1 (2)
(b)	$P(A' \mid B') = \dfrac{P(A' \cap B')}{P(B')}$ or $\dfrac{0.33}{0.55}$ $= \dfrac{3}{5}$ or 0.6	M1 A1 (2)
(c)	$P(B \cap C) = 0.45 \times 0.2$ $= \mathbf{0.09}$	M1 A1 (2)
(d)	 Allow 1st B1 for 3 intersecting circles in a box with zeros in the regions for $A \cap C$ Do not accept "blank" for zero	B1 B1ft B1 B1 (4)
(e)	$P(B \cup C)' = 0.22 + 0.22$ or $1 - [0.56]$ or $1 - [0.13 + 0.23 + 0.09 + 0.11]$ o.e. $= \mathbf{0.44}$	M1 A1 (2) **[12]**

Examiner reports

Question 1

Part (a) was answered very well and most candidates scored full marks here but responses to part (b) were mixed. Some thought that because both values were similar, but one positive and one negative, they "cancelled out" and others only commented on one of the tests or thought that the correlation coefficients were between the two tests. However a number of fully correct solutions were seen.

Question 2

There were many incorrect answers of 38 given to part (a) where candidates, presumably, thought the question wanted the value that 75% of candidates exceeded rather than the value exceeded by 75% of the candidates. They should be encouraged to read the question carefully!

Part (b) was answered very well and where they showed some calculations to determine the outliers they usually scored well on part (c) too. A surprisingly large number though failed to show any working for the outliers and their diagram showed no evidence of them either. There was still some uncertainty about where to end the whiskers (either at the outlier limits or at the next non-outlier value are acceptable) and there was some evidence that candidates had been using previous mark schemes as "model solutions" and were drawing both sets of whiskers which is, of course, incorrect.

There were many good answers to part (d) but some candidates failed to secure marks as they did not use the appropriate terms. The use of "mean" or "average" or "spread" are not sufficiently clear: we require use of the correct terms such as "median", "range" or "Inter Quartile Range" .

Question 3

Part (a) and part (b) were generally very well done with few candidates not knowing the correct structure of the tree diagram. A number did not fully label the tree diagram thus potentially losing the mark for the probabilities. Some candidates do not help themselves or the examiner by drawing very small diagrams. In part (b) it was pleasing to see very few candidates resorting to decimals and those who did seem to have got the message that exact equivalents are required using recurring decimals where appropriate. In part (c) many candidates demonstrated a lack of understanding of conditional probability. They could not transfer the context of the question into a formula and many still use P(A/B) with no explanation as to what A and B represent. Of those who did manage to write P(F'/L) many failed to see the significance of part (b)(ii).

Question 4

The first three parts were generally completed with confidence and fluency. Only a few candidates found incorrect values for the median (as they did $\frac{45}{2}$ and then looked for the 22.5th value) or the quartiles. The standard deviation was particularly well answered, but the usual errors of not dividing sum of x^2 by 45 or forgetting the square root were the most common.

In part (d) most candidates knew a rule to apply but a few got muddled with what exactly it was, using $Q_3 - Q_1$ or $Q_1 - Q_2$ or incorrectly for example. Having managed to apply the rule, some then got muddled in their reasoning, stating it showed positive skew or put the inequality signs the wrong way around which led to an incorrect conclusion. Those who simply used mean < median tended to fare better.

Part (e) was poorly answered, with some candidates not even attempting an answer. Perhaps they did not realise that this question considered a 'new' data set meaning it had little to do with previous work, with many performing calculations on the sample of 45 rather than considering the population. There were many scripts containing long and elaborate wrong calculations. The scaled mean was done better than the scaled standard deviation. On finding the new mean many candidates correctly subtracted 5 but multiplied their answer by 0.1 and did not subtract. A large number of those who attempted decoding applied the same rule to both the mean and standard deviation or stated incorrectly that the standard deviation was not affected by coding. Only a few candidates had intrinsic understanding that a measure of spread was only affected by the 'multiplier'. Those candidates who knew what to do produced short, efficient solutions.

Question 5

The use of interpolation, which was expected in part (a), is improving and many candidates made a reasonable attempt. Some failed to use correct class limits and had a width of 9 rather than 10 but the correct answer was often seen.

Despite the values for $\sum t$ and $\sum t^2$ being given in the question a number of candidates chose to estimate these values from the table and subsequently obtained an incorrect estimate for the mean but the mark scheme did permit them to obtain the remaining marks. The question did not ask for estimates of the mean and standard deviation and so using the mid-points was inappropriate in this question. The usual crop of errors arose when calculating the standard deviation with many failing to divide the 69378 by 32. Candidates can usually calculate S_{xx} correctly using the given formula and they may find it helpful to simply remember that standard deviation is simply $\sqrt{\frac{S_{xx}}{n}}$ but few seem to use this approach.

In part (c) most compared the mean and median and correctly concluded that the skewness was positive. Some used a formula which, apart from the extra work, was fine but those who used a (median – mean) formula rarely gave the correct conclusion. Some candidates went to great lengths to calculate the quartiles and conducted a quartile test for skewness which, if correct, led to a conclusion of negative skewness. This was allowed but was clearly not the intended approach for 2 marks.

Question 6

The vast majority of candidates produced accurate scatter diagrams and on the rare occasion that there was a point missing it was predominantly point *D*. Explaining exactly why a linear regression model was appropriate proved to be difficult for candidates overall. Most candidates seemed to have the general idea but did not express this in the required terms and consequently very few earned this mark. Comments tended to be much more general about why linear regression is carried out and most talked about correlation being high without explaining that the points lie close to a line.

On the whole the correct formulae were used in calculations of S_{dd} and $S_{fd,}$ with most candidates earning the method mark at the very least. The same was true in the calculations of *b* and *a* overall, although a common mistake was to calculate S_{ff} and go onto use that in the calculation of *b*. Premature approximation cost many candidates accuracy marks. Interpretations of the value of *b* were considerably varied, with relatively few candidates gaining this mark and some opted to omit this part altogether. Most candidates failed to relate their value to the context of the question and often tended to discuss *b* merely in terms of being the gradient. As a consequence, despite having the right kind of idea and correctly understanding the concept of the gradient, frequently candidates failed to gain this mark due to missing out the relevant units, mixing up the units or not quoting the actual value of *b*.

Very few candidates were able to formulate the correct equation with the correct units in part (f), and the majority found this particularly challenging, either omitting this part or resorting to evaluating the lines at the data points rather than equating and solving the equations. Often no clear strategy was apparent and a common mistake was to equate their equation to 5. There was clearly confusion over *t* and *d* and even out of those who were able to solve the required equation or inequality, not many found the value of *t* or range of *t* in km, as most tended to give their answer in terms of *d*. Occasionally the intersection point was evaluated using their graph after the lines had been plotted.

Question 7

Part (a) was answered very well with only a small minority adding $\mathrm{P}(A \cap B)$ rather than subtracting it. In part (b) although most could quote a correct formula, some thought that $\mathrm{P}(A' \cap B') = 1 - \mathrm{P}(A \cap B)$ and then had a numerator greater than their denominator. Others assumed independence and used $\mathrm{P}(A' \cap B') = \mathrm{P}(A') \times \mathrm{P}(B')$. Part (c) was usually correct although a small number of candidates confused independence with mutually exclusive and gave an answer of zero.

Most candidates gave a correct arrangement of the events in part (d) and few omitted the "box". There were a number of errors in the calculations of the other probabilities, in particular the probability for *B* only, but a large number of fully correct diagrams were seen. Q7(e) was usually answered quite well with many demonstrating a correct understanding of the method by using the values from their Venn diagram and others avoiding any problems with the Venn diagram by using $1 - \mathrm{P}(B \cup C) = 1 - [0.45 + 0.20 - 0.09]$ or occasionally $\mathrm{P}(B') \times \mathrm{P}(C')$.

Statistics for S1 Practice Paper Silver Level S1

Qu	Max Score	Modal score	Mean %	Mean score for students achieving grade: ALL	A*	A	B	C	D	E	U
1	7		79	5.53		5.99	5.57	5.24	4.83	4.38	3.29
2	11	10	74	8.10	10.09	9.85	9.06	8.38	7.69	7.00	5.27
3	11		73	8.03		9.89	8.68	7.87	7.13	6.50	5.15
4	13		70	9.12	11.49	10.65	9.31	8.54	7.74	7.09	5.22
5	7		66	4.60	6.35	5.89	4.93	4.21	3.35	2.64	1.66
6	14		62	8.65	11.47	10.33	9.25	8.64	8.13	7.48	5.58
7	12	12	66	7.96	11.35	10.49	8.65	6.80	5.37	4.52	3.36
	75		**69**	**51.99**		**63.09**	**55.45**	**49.68**	**44.24**	**39.61**	**29.53**