THE ROYAL STATISTICAL SOCIETY

ORDINARY CERTIFICATE
(2 papers)

SOLUTIONS 1998

Note:
Marks are given for neatness and clarity when constructing tables and diagrams.

Royal Statistical Society
ORIGINAL CERTIFICATE
May 1998
PAPER I

1.(i)Possible uses include:

Establishing Spending Patterns
Examination of inflation
Preparation of Retail Price Index
Regional information
UK national Accounts
National estimates of consumers' expenditure
Obtaining information on housing and transport
Aiding decisions on peusious and welfare benefits.

(ii)Stratified: the country is divided into strata(areas) of similar income, and/or similar region, and samples are chosen from these strata.

Clustered: more compact areas within each stratum are chosen for study, rather than taking households purely at random from the whole area.

Random: within the (randomly chosen) clusters, the households are sampled at random(in the UK, this is done using the Postcode Address File).

Visits need to be carried out at all seasons of the year to establish a reliable picture of expenditure.

(iii)The 14 categories actually used are:

Housing	*Fuel, Light and Power*	*Leisure Services*
Food	*Clothing and Footwear*	
Alcohol	*Household Goods*	*Personal Goods and Services*
Tobacco	*Household Services*	*Fares and other Travel costs*
Motoring	*Leisure Goods*	*Miscellaneous*

(iv)Sampling errors could arise because of the relative size of survey, which is about 7000 households in the UK, As compared with the total number of households.

Non-response bias could arise if households of a particular composition(eg single elderly persons), in a particular stratum(area or region) refused to take part or failed to complete a reliable diary of expenditure.

Incorrect reporting could include failure to record, e.g., alcohol or tobacco costs because current opinion regards these as "less worthy"; conversely items seen as "good" may be overestimated.

2.(i)Many schemes are possible, for example:

(1)Let the pairs	$00, 01$	correspond to ball number	1
	$02, 03$		2
	$04, 05$		3
	$\ldots$		$\ldots$
	$96, 97$		49
	$98, 99$	*be discarded.*	

(2)Discard 00,99 and use pairs (01,02), (03,04),$\cdots$ for 1, 2,$\cdots$, ending with (97,98) for 49.

(3)Let the pairs

$01, 51$	correspond to ball number	1
$02, 52$		2
$03, 53$		3
$\cdots$		$\cdots$
$49, 99$		49
$00, 50$	*be discarded.*	

(4)Use pairs module 49, i.e.(01,50)$\equiv$1; (02,51)$\equiv$2;$\cdots$(49,98)$\equiv$49, and discard 00,99.

Simulate the choice of 6 balls by the above methods, discarding any repeats and making extra choices until six distinct numbers have been found.(Method(4) is illustrated below)

(ii)

	(a)	(b)			(c)
12	12		89	40	
00	·		04		4
58	9		18		18
40	40		07		7
00	·		50		1
51	2		29		29
05	5		32		32
19	19		↑	3 *samples*	
26		26	*digits*	(a) (b) (c)	
74		25			
76		27			
57		8			
53		4			
57		(*repeat*)			

(iii)Samples are (a) 2 5 9 12 19 40 $-$ *none consecutive*

(b) 4 8 25 26 27 40 $-$ $25, 26, 27$

(c) 1 4 7 18 29 32 $-$ *none consecutive*

One of the three draws gives at least two consecutive numbers.

A computer could be used to generate a large number of sets of six, and the proportion of these with consecutive numbers could be used as an estimate of the probability.

3.(i)B is preferable because it is a random sampling method which allows sampling variation to be estimated; there is no selection bias; any estimates made from the 25 provide valid estimates for the population of 4000. It may be objected that B is much slower because random number tables must be used for the selection process, but method A excludes some companies altogether and does not give equal probabilities of selection for the others.

(ii)Stratification according to size of company is likely to be useful. For larger companies, with large numbers of transactions, there is more scope for error; perhaps the errors also will be larger and will have more effect on the resulting figures. Stratification allows errors to be estimated within strata, as well as the actual measurements, so providing useful information. This would improve method B by ensuring that all sizes of company were represented.

4.(i)(a)A sampling frame(e.g. constructed from the payroll) will be required, listing employees by departments. A random sample is then selected from each department(using methods such as those mentioned in earlier questions), and these particular employees are asked their views.

Departments may be sampled with probability proportional to size, or in any other convenient way (e.g. to minimize survey costs). The survey should be carried out as nearly as possible at the same time so as to minimize the effects of bias due to changes of mind through pressure group intervention etc.

(b)Only a list of departments is required. A quota(number to be sampled) is set for each department, and any available members of a department, up to the necessary quota, are surveyed. There is always the possibility that those most easily available will not be fully representative of the department; subdividing the department, e.g. by age, or length of employment, can be helpful, although then this extra information is required as well as a list of departments.

(ii)Total numbers of employees=2000. Required sample size=200. The numbers required from each stratum are therefore 1/10 of the totals in the departments, namely 15,25,5,130,5,20.

5.A pilot survey is intended to be a small-scale version of the eventual full survey, using the intended questionnaire and method of data collection.

It serves as a pre-test of the questions, examining whether they are acceptable, clear and understood by respondents; and of the ordering of questions, whether logical; whether sufficient alterative answers have been provided; whether respondents consider some questions irrelevant or personal, or the questionnaire too long. Open-ended questions may be tested in a pilot, with a view to providing sufficient boxes to tick in the final version.

The adequacy of the proposed sampling frame may be checked for completeness, accuracy and ease of use, the variability in the population estimated, so as to determine the required sample size, and the appropriates of the proposed data collection method examined; the relative accuracies and costs of possible alterative methods may be considered. Non-response rate may depend on the collection method, the need for re-visiting, the distance apart of the chosen units in the population etc.

When interviewers need special training, the success of this can be assessed from the results of a pilot; and finally the likely duration and cost of a full survey of the required size can be estimated, so that if necessary possible economies can be suggested. Efficiency and appropriateness of field organization may be judged.

6.(i)The sampling frame is all people watching the TV programme.

(ii)This sample is self-selecting; people require immediate access to a telephone(and to be fortunate that there is a line available when they choose to ring); double counting of groups or individuals who are determined to get through is possible; people who are interested enough to try will be likely to have a bias in favor of change.

(iii)Depending on the care of choice, or the success of finding a "representative" group, there may not be serious bias, although the need to reach a consensus may suppress or reduce natural variation. If any numerical output is needed(eg to recommend a level of investment or a purchase price), a small group like this may give an estimate which has poor precision.

7.(i)The form should include space for all of the following:

(a)interviewer identification, date and time of interview
(for record process and following up checks);

(b)school identification: name and postcode
(to identify the area in which the child lives);

(c)child identification:first name, sex

(first name for use in informal interview; sex in case there are major differences in expenditure);

(d)surname also requested for accurate identification;

(e)age,as years and months and date of birth(cross-check);

(f)list of all items upon which money may be spent, e.g. books, clothes, food and drink, entertainment, sports, travel, music, hifi etc. For use at home, discos, and anything else likely to be available where the survey is carried out(a pilot may have been done to get as complete a list as possible, but an " others" category with space to give more details should be included) with boxes to tick and space to record amount spent in the past week;

(g)amount of spending money available per week(either ask for exact figure or provide boxes giving ranges to tick);

(h)whether some or all is earned, and if so how(Saturday job, paper round etc.)or all is given as pocked money.

(ii)(a)interviewer identification field width 3 digits(or less); date six digits, time four digits;

(b)name first 8 letters; postcode 7 spaces;

(c)name 10 spaces; sex 1;

(d)10 spaces;

(e)4 spaces and 6 spaces; all digits;

(f)box for tick; 4 digits for sum of money(k.p);

(g)(h)similar to (f).

Aim to use only questions which give real information; arrange in a logical orders; number the questions; provide interviewer with detailed specification of what is to go in each category of (f), eg division between "entertainment" at home and elsewhere.

Establish whether a "clothes allowance" is given which includes essential school clothes, whether own money must be used for participation in organized sport, etc. A yes/no box with a single space(enter Y or N) should be adequate for each of these.

Answers appropriate to the country of residence will of course be accepted.

8.(a)Check that dd is between 01 and 28 for any values mm, 01 and 30 for mm=04,06,09,11, allow 29 for mm=02 when yyyy represents a leap year, range 01 to 31 for mm=01,03,05,07,08,10,12.

Check mm between 01 and 12. Check yyyy between 1990 and 1998(up to now).

Check that the admission date is earlier than the discharge date.

This may be done either by checking y,m,d in order or by computing the length of stay as the difference between the two dates(expressed in days) and check that this is positive(or, if day patients/outpatients are included, non-negative).

(b)After the checks above, 'freak' values should all be large positives, and there may be geriatric or psychiatric patients. Hence check classification of patient when following up 'freaks'. Such values, whether genuine or not, will increase means and variances, but have less effect on medians and quartiles. They will also cause greater skewness of data and so make comparisons between means less reliable. 'Freak' values maybe identified by constructing box-and-whisker plots, both with and without the suspect figures.

PAPER II

1.(i)Punches thrown in a World Heavyweight Championship Boxing match.

round	Number of punches thrown by		Total
	Tyson	Bruno	
1	43	55	98
2	39	42	81
3	28	35	63
4	37	18	55
5	55	20	75
	202	170	372

Source: Independent newspaper, 27 February 1998.

(ii)(a)202/5=40.4

(b)(55+42+35)/170=132/170=77.6%

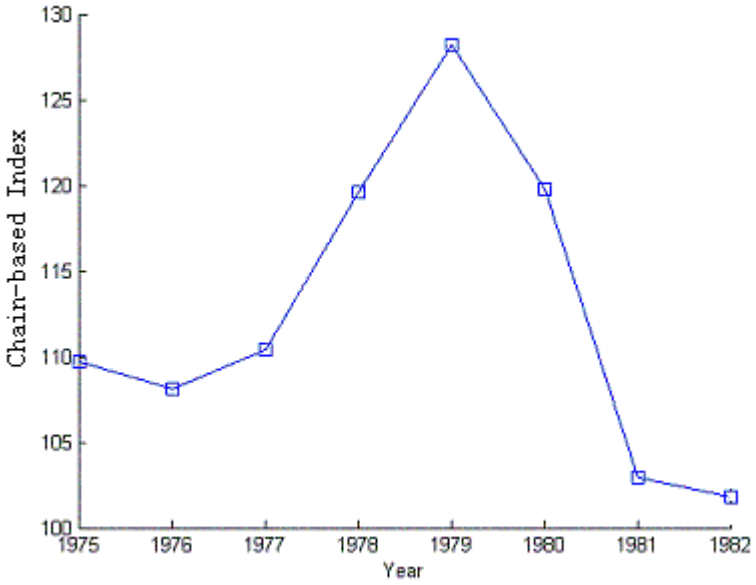
(c)55/75=73.3%

2.(a)(i)This is an index number where each period in the series is referred to the previous period as baseline(rather than a fixed past period).

(ii)The chain-based method can show more clearly the rates of change in the series, period by period, as well as the sizes of the changes over time. Rates of change are not easily seen by other methods.

(b)(i)A chain-based index is appropriate. Chain indices are:

1975	1976	1977	1978	1979	1980	1981	1982
109.7	108.1	110.4	119.6	128.2	119.8	102.9	101.8



(ii)Over the whole period, the increase from 1975 was 152.2%, because $\frac{28500}{11300}$=2.522. This

represents about 12.3% P.a. on average, since $(1.123)^8$=2.530.

There was a rise every year, since the chain index is always above 100. Up to 1977, there was about 10% rise each year, followed by two very sharp rises and then two large falls, with the index at the end of the period below 2% rise. The greatest rise was for 1979, namely 28%.

3.(a)(i)When data are arranged in increasing order of size(ranked) the p^m percentile is the value below which p% of the observations lie.

(ii)Rank the 50 observations. Then we require the p^m and $(p+1)^{th}$ values. It is usual to place the p^m percentile p% of the distance between the p^m and $(p+1)^{th}$ observations, so that the 10^m percentile is found as $0.1\times p^m$ value + $0.9\times(p+1)^{th}$ value.

[sometimes it is taken midway between p^m and $(p+1)^m$.]

(b)Ages in rank orders:

$$13,16,33,34,35,41,42,43,\underbrace{43,43}_{lower\ \ quatile},47,48,49,49,50,50,51,53,\underbrace{55,56}_{median},56,56,59,60,$$

$$65,65,67,67,\underbrace{67,68}_{upper\ \ quartile},68,68,69,70,71,77,77,81,81$$

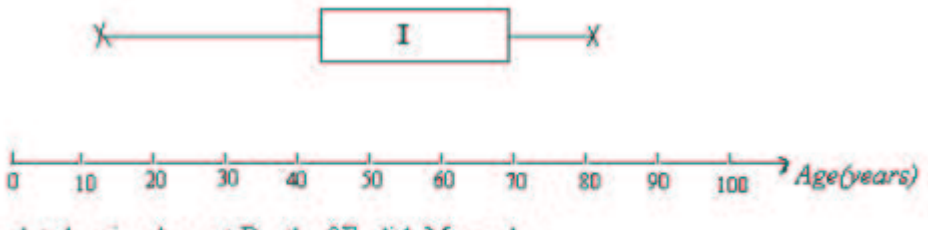
(i)N=40 observations.

25^M percentile(lower quartile) between 10^M and 11^M, both of which are 43, so its value is 43.

50^M percentile(median) midway between 20^M and 21^{sr}, i.e. $\frac{1}{2}$(55+56)=55.5.

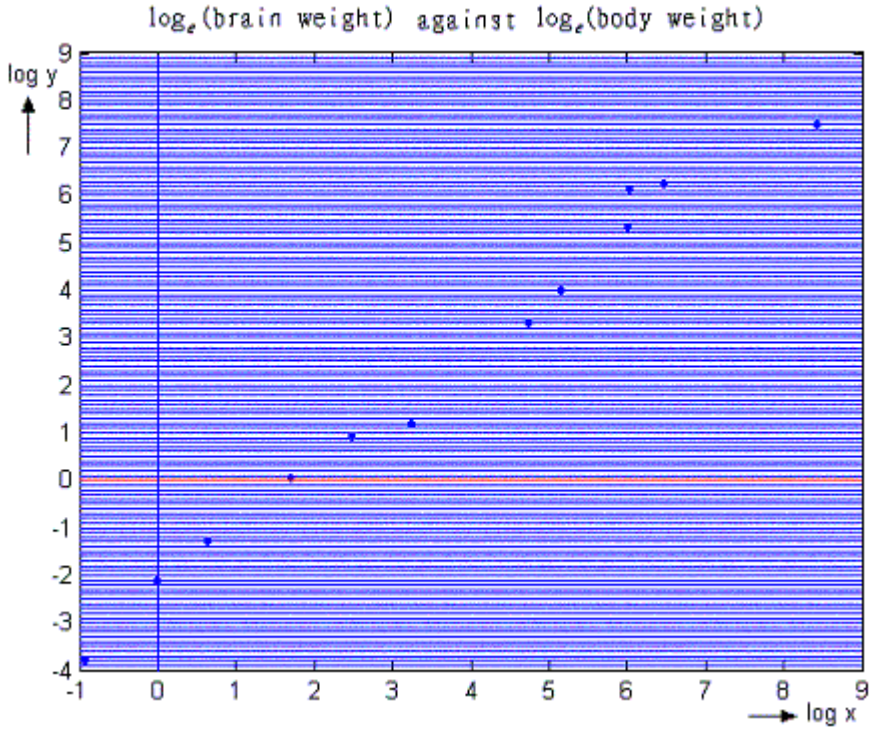
75^M percentile(upper quartile) between 30^M and 31^{sr}, which are 67,68; by the method given above it is $\frac{3}{4}\times 67+\frac{1}{4}\times 68=67.25$(or alternatively it may be taken as 67.5)

(ii)



Boxplot showing Ages at Death of English Monarchs.
[Note: median and extreme value can be shown with any convenient symbol]

4.(i)



(ii)Using r=0.994, B=t·$\sigma_y/\sigma_x = 0.994 \times 3.714/2.927 = 1.261$

$[t = \frac{S_{xy}}{\sigma_x \sigma_y}$ and $B = \frac{S_{xy}}{\sigma_x^2}]$

$\log_e A = \overline{Y} - B\overline{X} = 2.325 - 1.261 \times 3.671 = -2.304$

$[Y \equiv \log_e y; \quad X \equiv \log_e x]$

Hence $A = e^{-2.304} = 0.100$.

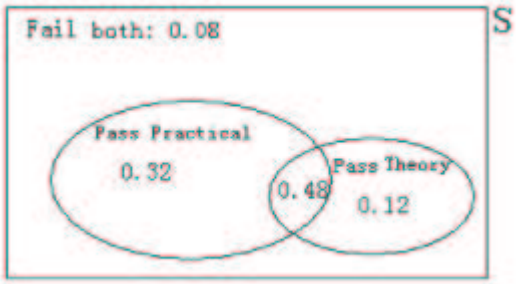
5(i)For Mr Smith, P(pass practical)=0.8, P(fail practical)=0.2

and so P(fail theory)=0.4, P(pass theory)=0.6).

P(pass both)=0.8×0.6=0.48 assuming independence.

Hence P(pass practical, fail theory)=0.8×0.4=0.32 and P(fail practical, pass theory)=0.2×0.6=0.12.

The events in the diagram are passing the tests, and the whole sample space consists of all four combinations of the results(total 1).



(ii)For Mr Jones, P(pass practical)=0.7, P(fail practical)=0.3, hence P(fail theory)=0.3 and P(pass theory)=0.7.

Assuming independence, P(pass both)=0.7×0.7=0.49.

He is more likely to pass than Mr Smith.

6. N=170. Measurements to nearest minute, so "40-49" includes every measurement between 39.5 and $49.5^{(-)}$, giving 44.5 as midpoint.

$Miapoint(x)$	f	fx	fx^2
44.5	9	400.5	17822.25
54.5	34	1853.0	100988.50
64.5	19	1225.5	79044.75
74.5	53	3948.5	294163.25
84.5	46	3887.0	328451.50
94.5	8	756.0	71442.00
104.5	1	104.5	10920.25
	170	12175.0	902832.50

mean=12175/170=71.6 mins.

(Coding may be used if desired; but not essential with pocket calculator).

Variance=$\frac{1}{169}$(902832.50-$\frac{12175.0^2}{170}$)=182.7671 and standard deviation = $\sqrt{182.7671}$ = 13.5 mins.

7.(a)(i)An additive model is appropriate when the change attributable to season is not dependent on the value of the series in each period; e.g. it can be assumed that quarterly effects are the same for every year.

(ii)A multiplicative model is used when the change due to season is an amount that is proportional to the value of the series in that season; the proportion will be the same each year.

Trend is removed by the Moving Average. From 1987(3) to 1989(2) we have:

Quarter	1	2	3	4
1987			1.49	−0.62
1988	−1.18	0.36	1.37	−0.61
1989	−1.02	0.42		
Average	−1.10	0.39	1.43	−0.615
Adjusted to zero (−0.03)	−1.13	0.36	1.40	−0.64

as the detrended series. $Total + 0.105$
$Total - 0.01$ *approx.*

These give the seasonal component.

(b)When data are not seasonal, it may be useful to give greater weight to more recent observations, especially for forecasting purposes.

Weighting can also be used to overcome missing values in the series.

If an 'outlier' is present, or there was some external influence likely to have distorted one or more observations, weighting can help to reduce the influence of these disturbance.

8.(i)A nominal variable does not take numerical values, but only has 'values' which are categories-e.g. male/female; opinions expressed as yes/no, for/against; characteristics such as eye color.

Ordinal variables can be expressed in some natural order, e.g. by categorizing, opinions as 'strongly disagree'with some statement, through to 'strongly agree', with perhaps another three intermedian categories like 'disagree', 'neutral', 'agree'.

(ii)A compound bar chart is a possibility-a multiple bar chart would also serve.

If either A or B can be regarded as a "response variable"it should be plotted in the y-direction. If neither is naturally a response(or if both can be so regarded) a choice is made which helps letter interpretation.

Suppose that A is the percentage of readers of different newspapers in a population which is subdivided by "social class" B.

For each class, there is a bar of constant height, representing 100%, subdivided according to the distribution of readership in that class. Different shading or coloring is needed to clarify the differences between these distributions in different classes.

In this example, there is a natural order within class labels as used in the UK-A,B,$\cdots$. This is used in the x-direction.

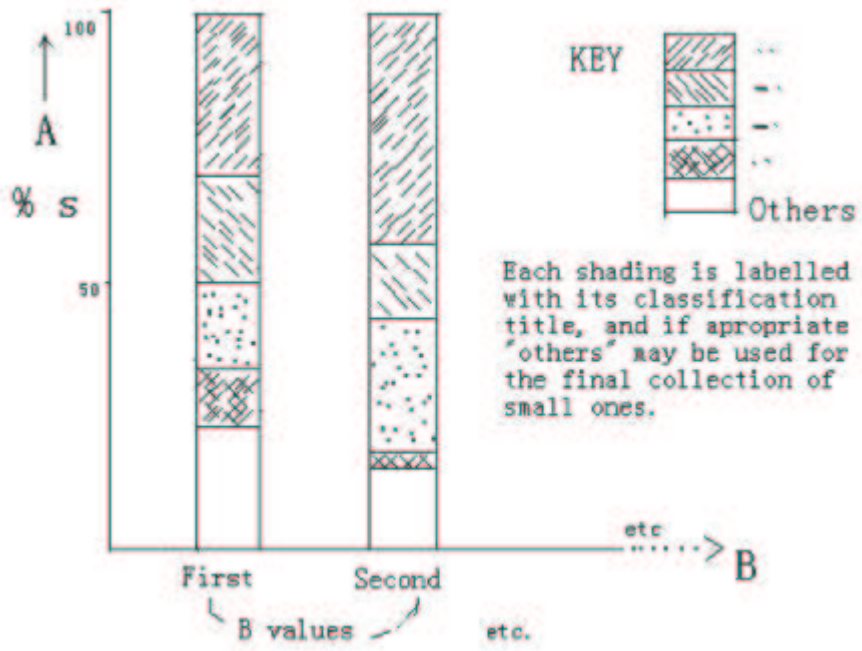
If there is no such natural order, one of the items of A can be chosen, e.g. that whose proportion shows sufficient changes through to different categories of B to be made the basis of a systematic plot in increasing/decreasing order.

When actual numbers are to be shown rather than percentages, the complete bar for A represents a total, and then B can, if appropriate, be taken in increasing/decreasing order of totals.

Of course if B represents time(eg successive years) this will normally be used in order of years/periods along the x-axis. The pattern for A will then vary in a way which may be less easy to interpret.

Clear labelling of axes is essential, and so is a key for colors or shading.

Sometimes actual numbers or percentages can be written inside the bars or sections of bars to



help interpretation.

A multiple bar-chart has this form: