The Royal Statistical Society

ORDINARY CERTIFICATE

IN STATISTICS 2000

SOLUTIONS

PaperI

1: (i) Possible questions:
Whether members of the household work
if so, what occupation and whether full-time or part-time
number of vehicles (all type) in household
type of accommodation, owner or rented etc, time at that address
if working, whether public transport is available to travel to work
socio-economic classification of household

(ii)Advantages:
response rate likely to be high
interviewer can clarify questions if necessary
interviewer can ensure complete information is given
questions can be less simple than when there is no personal contact

Disadvantages:
usually expensive, needs return visit if respondent not available
interviewers may show bias in asking or interpreting questions
interviewers may make recording errors, or show bias in deciding which box to tick for reply

(iii)Each month choose the 20 primary sampling units (which in the uk might be, for example, Postal Districts) to give good coverage of socio-economic conditions and rural-urban split,as well as geographically through the country. within each chosen primary unit, choose at random 210 addresses (e.g.using the Postcode Address File). Each interviewer can be given a relatively small and compact area by this method, reducing cost. Household in different types of area will have been covered in this way. Monthly sampling allows for the likely difference due to holiday periods,weather conditions etc. affecting people's travel patterns.

(iv) Assuring that new samples were chosen each month (as they should be) the total number of addresses aimed for is $12 \times 21 \times 20 = 5040$ of these only $5040 - 623 = 4417$ were eligible,but 58 contained two households, giving$4417 + 58 = 4475$ households.

Ineligible addresses may be employ at the time, or may be business premises or shops, hotel etc.[The extra households may live in flats, annexes or separate parts of a building: if a shop has a flat above it the occupants of that are included in the sample]

(v) Total is 4475, so D contains $4475-(3214+300+827)=134$ and the percentages are:

A	B	C	D
71.8	6.7	18.5	3.0

2.(i) There are altogether 10,000 patients, and 200 are required in the sample; this is a proportion $\frac{200}{10000}=\frac{1}{50}$of the whole list.Each doctor should provide $\frac{1}{50}$of his list ,i.e. A,50;B,60;C,30;D,60. The doctors should be given instruction how to use random number tables to select the right number of patients from lists in which each patient has been given a number (between 0001 and 2500 for A,etc.).

(ii) The same sample size could be used, but the doctor would be free to select any 50/60/30/60/of their patients. They might choose those who visited the medical center while the survey was given on, or they might subjectively select people from their lists and replace with another any patient who does not respond.

(iii)(a) Advantage:every patient has the same chance of being selected .

Disadvantage: selecting the sample requires time, and some technical knowledge of random sampling.

Also-advantages: sampling variation can be estimated, for each doctor and for the whole center; non-respondents can be identified;

-disadvantages: out-of-date lists cause non-response;seriously ill patients are not able to answer; those samples may not have visited the center recently and do not know much about it.

(b)Advantage:no non-response problem,because a replacement is used .

Disadvantage: no estimation of sampling error variation.

Also-advantage: easy sample selection, no skill required; choice can be made from those who do use the center; speed of completing survey;

-disadvantages: no guarantee of sex, age, residential area being represented in right proportions; no information on non-respondents; patients probability of selection not constant.

3.(a)(i) $8/32=1/4=0.25$;(ii)$24/32=3/4=0.75$ Label children 1-32. Associate pairs of digits 01,33,65 with child 1; 02,34,66 with child 2; up to 32,64,96 with child 32.Do not use the pairs 00,97,98,99. Now each child has the same probability of selection in the sample.

digirs :	96	30	28	60	95	32	96	31	04	88	86	44
child :	32	30	28	28	31	32	32	31	4	24	21	12
	√	√	√	×	√	×	×	×	√	√	√	√

Discard any repeat selections×
The resulting samples is 4,12,21,24,28,30,31,32.
Note: if 00 and all numbers from 33 upwards are discarded,this gives 4,17,19,25,28,30,31,32,but is much more wasteful of digits as discards.

(b)Number the books 000-999, so that they correspond with categories as follows:

General	*Crime*	*Romance*	*Horror*	*western*
$000 - 431$	$432 - 742$	$743 - 854$	$855 - 943$	$944 - 999$

[ALTERNATIVELY,start at 001 and use 00 to stand for 1000, so that all the above boundaries move one number up]
select 8 books, as below,and place them in the categories that are given by their 3-digit numbers.

856	074	571	965	640	332	410	485
Horror	*General*	*crime*	*Westen*	*Crime*	*Genreal*	*General*	*Crime*

No sets of digits needed to be discarded because they were repeated

4.Advantages:Panel method usually gives greater precision than separate individual samples of the same size; it is useful for studying the effects of particular events or measures introduced at specific times ; people who change their view can be studied; it is easy to administer in practice; it can produce fuller and more reliable data.

Disadvantage: People may not be willing to take part in repeated surveys, so initial recruitment may be restricted; there are likely to be some losses from the panel as time goes on ; panel members may be become ”conditioned“ to the questions and cease to be representative of the wider-population.

5.(i) although English is the standard language for airline operations ,not all passengers may understand it, so useless copies are provided in other languages (e.g. languages of points on the route where passengers may get on or off) there will be non-response.

Using pre-selected seat numbers may lead to children being asked to complete the questionnaire ,or to empty seats and if these are not replaced by 'reserves' the sample size is reduced(parents of accompanied children could be asked instead.) some stratified or quota method applied to those who actually board the flight may be instead.

Issue and collection by cabin crew could lead to bias; the crew know who have the questionnaires and may take more care of them; passengers may not wish to grumble at those who have served them. seat numbers should not be shown on the questionnaires and ideally someone other than cabin crew should distribute and collect the questionnaires.

(ii)(1) simplify to obtain unambiguous answers for 0700 and 2200 ;

Was your flight DAY(departure between 0700 and 2200hrs) or NIGHT(after 2200hrs and before 0700hrs)?
□ Yes
□ No

(2)Respondents are unlike to know the conditions at the start of the flight,and can not give a useful answer if they visited once only or not at all. simple wording and more direct questions could be used :

were toilets clean and tidy whenever you visited them?
□ Yes
□ No
□ Did not visit

If your answer is no, at what stage of the flight were they not satisfactory?
□ Beginning
□ Middle
□ End

(3) The question is ambiguous because people may not flight at all during the six months, or the airline may not operate the route they wish to use the next. a more precise line of questioning is:

Are you intending to fly again during the next 6 months?
□ Yes
□ No

if yes ,will you use ABC?
□ Yes
□ No

if you intend to fly with another airline,why?
□ ABC do not fly that routes
□ Prefer another airline

It would be possible to go on and ask about price, and/or quality of service, convenience of schedule etc.

6.(i) NAME:Family name(surname)______
First name(s)_____
PRESENT HOME ADDRESS:___________POSTCODE_____
TELEPHONE____FAX _____ EMAIL____
YEAR OF GRADUATION____ TITLE OF DEGREE_____

ANY FURTHER QUALIFICATIONS OBTAINED:

(1)Dgrees Level(egMA) Name of institution Year Subject or topic
1.
2.

(2) Professional Qualification Vocational Training
Title Year Awarding Body/Institution
1
2

Employment: Give present employer first,others in date order
Employer's name & Address Dates Title of post held Fulltime job or part time
1
2
3
If self-employed or a consultant,please say in this column,and give other details.
Present employer
Telephone number of your line manager or department head

____
FAX_____________
EMAIL______________

(ii) lack of up-to-date contact addresses; people not interested in replying;unwilling to give some of the details; abroad or hard to contact through inability or type work; Those who do reply may be in contact with others; send them a list of non-responders. Use any remaining contact with relevant academic staff to assist with updating address lists. use professional registers(societies, professional bodies)in relevant subjects.Advertise on website or local and national press. Include an encouraging covering letter urging people to reply and saying how useful a good database will be to the university.

7. Time and motion, if carried out unobtrusive, allows the exact working of a department to be studied in detail,without replying on biased or incomplete verbal information. various parts of a job can be timed, and interaction among employees worked. there is no interference with on going work. Interviews can discover in addition the views of employees, e.g.an possible improvement or reorganization of operations are done in a particular way, what else has been tried and either change or discarded.

8.

$FIELD\ \ NAME$	$FIELD\ \ TYPE$	$WIDT+1$	$DECIMALS$
SKU	$Character$	7	
$Irem$	$character$	40	
$Cost$	$numeric$	7	2
$sell-price$	$numeric$	7	2
$stack-quan$	$numeric$	6	0

paper II

1.(i)If the mean and standard deviation of the set of value are $\overline{x}$, s. the % coefficient of variation is $\frac{100s}{\overline{x}}$. It is useful as a dimensionless measure of variability in the data relative to their general size (i.e.it does not depend on the units of measurement).
It can be misleading in cases where measurements may take negative values, and is unstable if $\overline{x}$ is near to zero. If the origins of two scales of measurement, such as ^{o}C ^{o}F for temperature, are different, the %CV's on two scales will also be different.
sometimes important information may be lost by not quoting both $\overline{x}$ and s separately.

(ii) Maximum is 32.65, minimum is 32.55
For the standard deviations,$max = 3.05$,$min = 2.95$
Maximum for %cv is $\frac{100\times3.05}{32.55} = 9.4$
and minimum is $\frac{100\times2.95}{32.65} = 9.0$

2.(i)$Total = 20 \times 3.06 = 736$; hence the connected total is $736 - 34.5 + 45.3 = 746.8$ and the mean is 37.34

(ii)If the uncorrected sum of squares is $\sum x^2$, then $s^2 = \frac{1}{19}\{\sum x^2 - \frac{[\sum x]^2}{20}\}$
so $19 \times 2.90^2 + \frac{736^2}{20} = \sum x^2 = 159.79 + 27084.80 = 27244.59$

(iii)The corrected $\sum x^2 = 27244.59 + 45.3^2 - 34.5^2 = 2810643$
The corrected total is 746.8
Hence $s^2 = \frac{1}{19}\{28106.43 - \frac{746.8^2}{20}\} = \frac{220.918}{19} = 11.6272$ and $s = 3.41$.

(iv) Both have increase: the uses value is considerably larger than the one it replaced. so the mean rises; and the new value is also farther form the mean that was previously the case, so that standard deviation increases.

3(i)21M,   39F.

(ii)

RATING	1	2	3	4	5	*TOTAL*
FREQUENCY	12	17	13	8	10	60

(iii)

RATING	1	2	3	4	5	*TOTAL*
MALE	7	5	9	0	0	21
FEMALE	5	12	4	8	10	39
	12	17	13	8	10	60

(iv)Percentage by rows

RATING	1	2	3	4	5
MALE	33	24	43	0	0
FEMALE	13	31	0	20	26

(e.g.33%is 7/21,expressed to the nearest whole number)
Although the overall table in (i) showed a reasonably well-spread use of categories,we see from this percentage table that in fact all the 18 who disagreed with the statement were female.About half the males were neutral, but few of the females. These difference between the sexes may be explained by different size of different time of use for different purposes(shopping may not be the only use), but without more information we can only note the results. “Easy to manoeuver”is by no means proved.

4(i)(a)The Laspeyres index is $\frac{100\sum p_0q_0}{\sum p_0q_0}$, where$p_0$ p_1 are price in 1995, 1999 and $q_0$is quantity used in 1995.

$$L = 100\times\frac{(5\times 42.00)+(10\times 13.99)+(10\times 4.29)+(15\times 0.72)}{(5\times 22.49)+(10\times 8.69)+(10\times 3.99)+(15\times 0.57)} = \frac{100\times 403.60}{247.80} = 162.9$$

(b)The Paasche index is $\frac{100\sum p_1q_1}{\sum p_0q_1}$, where q_1 is quantity used in 1999.

$$P = 100\times\frac{(20\times 42.00)+(30\times 13.99)+(10\times 4.29)+(10\times 0.72)}{(20\times 22.49)+(30\times 8.69)+(10\times 3.99)+(10\times 0.57)} = \frac{100\times 1309.80}{756.10} = 173.2$$

(ii)L shows a 62.9% increase in stationary prices over the period, whereas P shows a 73.2% increase.

L reflect price change if the same quantities were being bought, and P reflects the change if the new(1999)quantities had been bought in 1955. Paper product have gave up in price more than folders and paperclips. As the business developed it required of the more expensive items. L use the out-of-date-purchasing patten and so underestimates the percentage increase. P may overestimate the increase but it is better to use this for budgeting purposes. If a single index is required,$\sqrt{L \times P} = 168.0$ is useful and gives a fair indication of change.

5(i)All entries in this table have the same probability:there are 36 of them. $P(x < 4)$is those marked$\surd$, $\times$ which are 18 of the 36 so $P(x < 4) = \frac{18}{36} = \frac{1}{2}$

(ii)$P(x < 4 \ \& \ y < 3)$ is $\times$ and so $\frac{6}{36} = \frac{1}{6}$.

(iii)$x < 4ory < 3$ contains 24 of the 36 ,so $P(x < 4 \ or \ y < 3) = \frac{24}{36} = \frac{2}{3}$

(iv)only (x,y)=(3,1),(2,3),(1,5)give 2x+y=7,so $P(2x + y = 7) = \frac{3}{36} = \frac{1}{12}$

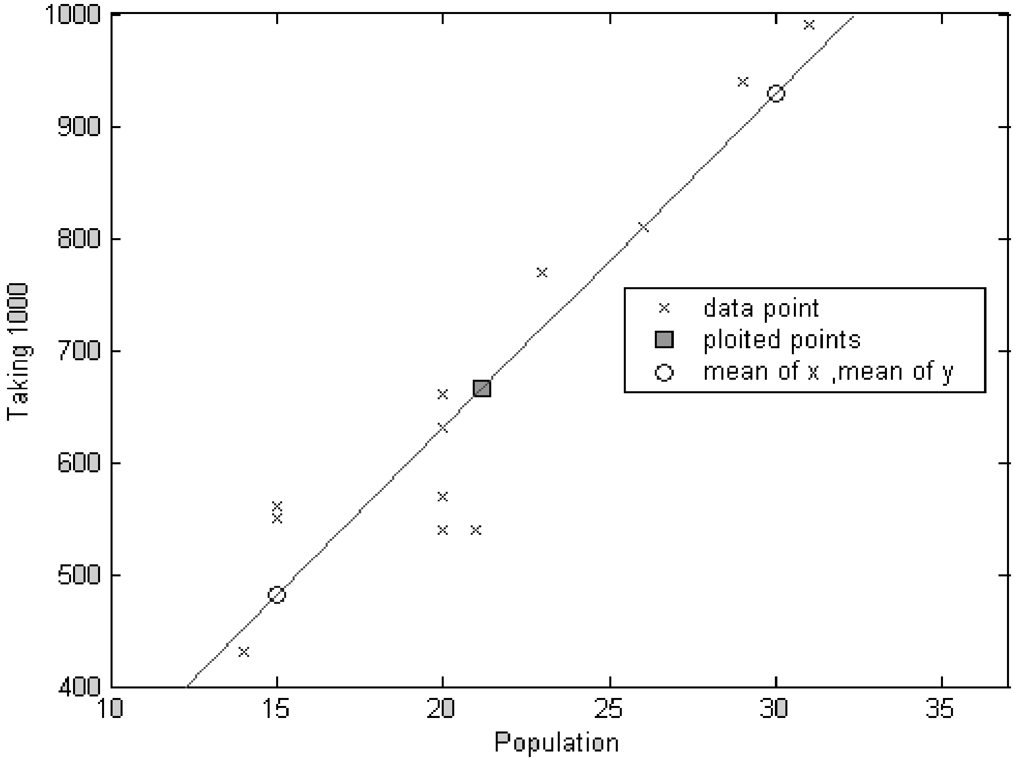
(v) $x < y$ for $\begin{array}{lll} x = 1 & y = 2, 3, 4, 5or6 & (5ways) \\ x = 2 & y = 3, 4, 5, 6 & (4ways) \\ x = 3 & y = 4, 5, 6 & (3ways) \\ x = 4 & y = 5or6 & and \ \ x = 5 \ \ y = 6 \end{array}$

Thus $p(x < y) = \frac{15}{36} = \frac{5}{12}$

(vi)The maximum of x and y is 4 for (4,1) (4,2) (4,3) (4,4) (1,4) (2,4) (3,4)i.e. 7 ways.

so$P(\max(x, y) = 4) = \frac{7}{36}$

6(i)As shown in figure1



The 6th question

Points lie reasonably close to a straight line with positive slope
A strong positive correlation is to be expected.

(ii)

$$r = \frac{178330 - \frac{(254)\times(7970)}{12}}{\sqrt{(5946 - \frac{254^2}{12}) \times (5629500 - \frac{7970^2}{1.2})}} = \frac{9631.6667}{\sqrt{317.6667 \times 336091.6667}} = 0.932$$

On 10 d.f., this is clearly very significant; stores with great local populations have larger takings.

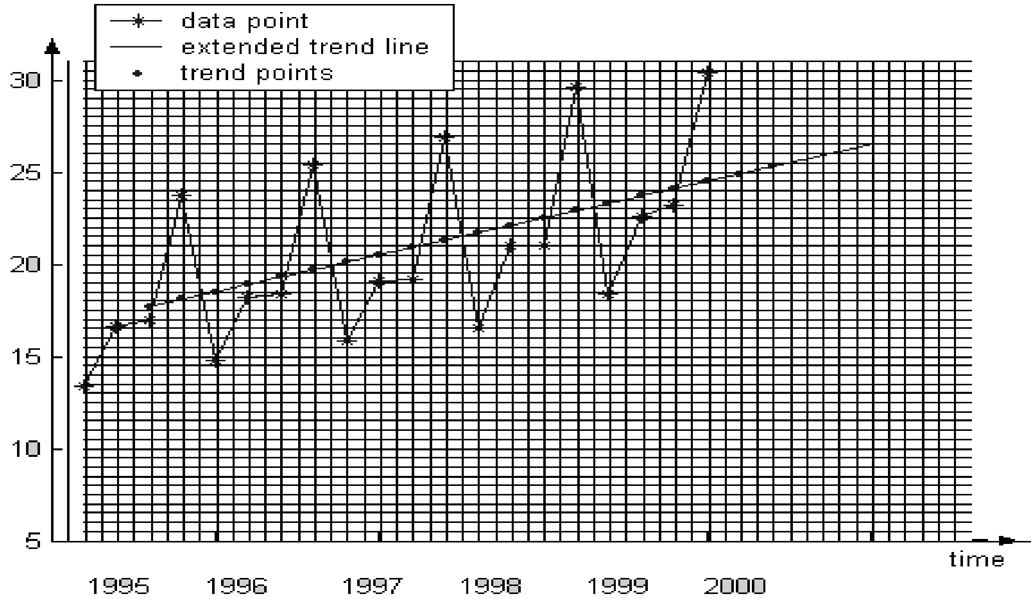
(iii) $y = a + bx$,where b=30.32(given) and $a = \bar{y} - b\bar{x}$.
Hence $a = \frac{7970}{12} - 30.32 \times \frac{254}{12} = 22.39$
we need two points to plot the line; suppose x=15 x=30. For $x = 15, y = 22.39 + 15 \times 30.32 = 477$,and for$x = 30, y = 22.39 + 30 \times 30.32 = 932$.(see points $\odot$ on graph)thought which the line is drawn. [It also passes through $(\bar{x}, \bar{y})$]

(iv)For x=22 y=689 and for x=35,y=1084
For x=22 the average taking of stores is predicted as £689000 but there is variation about this for individual stores (as for those with x=20 in the data set)

For x=35, we are outside the range of available data and although we may predict the average as £1084000 we do not have too much confidence in it because we can not be sure the linear relation still holds.(For example, larger populations may have greater choice of stores).

7(i)(a) Trend is the basic long-term movement of the series.

(b) The seasonal component is the short-term regular variation about the trend.It may occur over season of the year,but also days of the week or times of the day.



The graph shows a large but regular seasonal component,and a method based on moving averages is likely to do well in these condition

(ii)

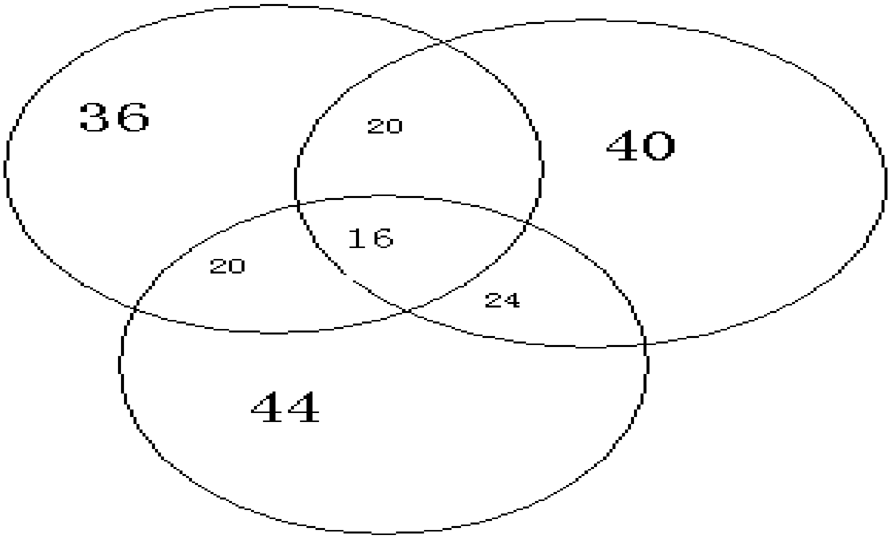
$Year$	$Quarter$	$sale(£)$	$4 - quartermovingaverrage$	$Trend$
1995	1	13.4		
	2	16.6		
			17.675	
	3	17.0		17.850
			18.025	
	4	23.7		18.225
			18.425	
1996	1	14.8		18.600
			18.775	
	2	18.2		18.988
			19.2000	
	3	18.4		19.325
			19.450	
	4	25.4		19.563
			19.675	
1997	1	15.8		19.775
			19.875	
	2	19.1		20.063
			20.250	
	3	19.2		20.388
			20.425	
	4	26.9		20.663
			20.900	
1998	1	16.5		21.125
			21.350	
	2	21.0		21.688
			22.025	
	3	21.0		22.263
			22.500	
	4	29.6		22.700
			22.900	
1999	1	18.4		23.175
			23.450	
	2	22.6		23.660
			23.650	
	3	23.2		
	4	30.4		

Extension to 2000 are, approximately:

$Q1$	25.0
2	25.5
3	25.9
4	26.4

[Note: over the whole period, the average quarterly trend increase is $\frac{23.5-17.85}{15} = 0.38$; but in the last two years the increase was a little more than this, and extending the final trend line allows for this.]
We have to assume there will be no sudden change in trend,due to economic or other factors.

8(i)As shown in figure3



O outside W F and P

$$\begin{array}{c}W\ and\ F - W\quad F\ and\ P = 20 = W\quad F\ butnot\ P\\ F\ and\ P - W\quad F\ and\ P = 24 = F\quad P\ butnot\ W\\ W\ and\ P - W\quad F\ and\ P = 20 = W\quad P\ butnot\ F\end{array}$$

After this step, use F=100 to find the figure for Falone,40;also use W=92 to give Walone=36

we know there were 200 accidents and W,F,P had some part in all of these,so Palone must be 44 to complete this total.

P is involved in 20+16+24+44=104 accidents.

(ii)One factor : W or F or P only :$36 + 40 + 44 = 120$ so$P(1 factors) = \frac{120}{200} = 0.6$
Two factors: $WF + FP + WP : 20 + 24 + 20 = 64$ so $P(2 factors) = \frac{64}{200} = 0.32$.
All Three factors, WFP=16, so $P(3 factors) = \frac{16}{200} = 0.08$.