$Ordinary \quad Certificate \quad in \quad Statistics$

$May 1999$

$SOLUTIONS$

Paper I

1 (i) The main use of the UK Index is to measure the rate of inflation. The main users are economists, business people, media, politicians and wage negotiator .

(ii) The prices of particulars goods such as food items are collected monthly by researchers going to a sample of shops in various parts of the country and recording prices on a fixed day each month. Prices for service items,such as hairdressing, are collected in a similar way. An averaging process is then used. Prices of items like postage, licence fees, travel fees are collected centrally from the relevant authorities or organizations. Weights are determined from a household expenditure survey, such as the UK Family Expenditure Survey which uses a sample of about 7000 households who keep records of all expenditure over a period of two weeks. From these records the average relative importance of expenditure in specific categories or components of the index can be calculated. The weights used are updated each years.

(iii) (a) All recent indices have been based on the index calculated in January 1987,which was assigned the value 100 so that each taker index is expresses as a percentage, 100x index calculated in January 1987.

(b) $\%index = \frac{154.4-150.2}{150.2} \times 100 = 2.8\%$

2 (i) Advantages: personal contact makes response more likely, clarifies questions, where necessary, ensures answers recorded property.
Disadvantages: cost and time taken, and possible interviewer bias.

(ii) Advantages: may make people more willing to take part in the survey and spend a little time giving useful answers.
Disadvantages: may make people join in just for the reward, without taking care to give true or relevant answers or having any interest in the survey topic.

(iii) Advantages: gives a basis for examining trends over time, may make interviewing easier as these is familiarity with the material.
Disadvantages: carries on any bias in the original sample, may lead to lack of participation because people are not willing to spend more time on it, may not be able to obtain all of them because of moving etc.

(iv) Advantages: large sample sizes; allows check of whether returns have special characteristics.
Disadvantages: cost and inconvenience of adding to database eta late stage, and delay in reporting if a strict cut-off date is not kept to.

3 (i) Choice of variables should be those that are strictly relevant to the study. Form should be sample to encourage accurate completion some entries, e.g. reference number, assume access to polices records.

Accident reference number ____

Name, address, position(police officer, medical/ambulance worker, etc.)of person completing report ____

Data and Time of Accident:

Data ____________________ Time ____________________

Location of Accident:

Number of road ___________

Location Identifier: grid reference, description of position in roms of distance from identifying feature/landmark ____

Tick appropriate boxes:

Bend in road	YES NO	*cross − roads*	YES NO
T − Jonction	YES NO	*Visibility obscured*	YES NO
Road Works	YES NO	*Incline*	YES NO
Pedestrian Crossing	YES NO	*Traffic Lights*	YES NO

Vehicles Involved. List AGE, MAKE, MODEL, cc for each___________

Number of vehicles involved and assessment of damage.

Type of Vehicle	Number involved Total	Severe	Moderate	Minor	None
Heavy Goods Vehicle					
Van					
Public Service Vehicle					
Car					
Motor cycle/mobed					
Bicycle					
Other(please specify)					

People Involved: enter NUMBERS in appropriate boxes.

Driver [1] Passengers □ Motorcyclist/passager □ Cyclist [0/1] (cross out one)

Pedestrians □

Driver Sex(Tick box)

M	
F	

ESTIMATED AGE (Tick box)

under 30	30-60	over 60

PRESENCE OF ALCOHOL

YES
NO

Injures: TOTAL NUMBER _________

Details of each:	fatality	serious-hospital case	minor-not hospital case	none
drivers				
list other passengers				
others				

Weather Conditions:

FOG	YES	RIAN	YES	SNOW	YES	ICE	YES
	NO		NO		NO		NO

Indicate severity of each of these (below) where relevant.

_________ _________ _________ _________

Road Conditions:

MUD	YES	SPILLAGE(oil,etc.)	YES	SURFACE DAMAGED	YES
	NO		NO		NO

ROAD: Street Lights on

YES
NO

or no Street Lights □ .

VEHICLES:(List)

HEADLIGHTS		SIDE LIGHTS	FOG LIGHTS	NONE
FULL	DIPPED			

Police Information. DRIVER-previous motoring offences(last 5 years)__________ previous accidents(last 5 years)__________ length of time driving(years)__________ presence of drugs YES/NO__________

(ii) Time can be given to nearest minute using 24 hours clock, with 25.00 as the missing value code. Cross road be 1(YES) or 2(NO)with, say, 9(or0)as missing value code.

4 (i) Good points include :
Clear instructions on filling in and returning the questionnaire, use of pre-coded questions, e.g.1(question3), clear format layout for giving date of birth, thanks for answering, offer of incentive for taking part, provision of boxes to tick, to simplify answer, classifying income by ranges, not asking for an"exact"number, in past 1, question 4 gives a cross-check for question 5.

(ii) Bad point include :
overall, the questionnaire has too many numbers used for different purposes-instructions, parts, individual questions, boxes in part 2, question 5 relies on a long memory period,

question 4 does not specify packet sizes, and is over too short a time scale;monthly better man weekly, what are "children cereals"?, question 3 does not state any time out-per week? /visit?/ $\cdots$, question 2 could be distance or time; should specify, in part 1, question 5, how do we deal with twins of the same sex?, some people will not give their exact date of birth-although they may still object if boxes were provided, the personal age/incame questions come too in the questionnaire, 6-16 weeks is a very long delivery time for coupons-no real incentive.

5 (i) For proportional allocation of a total sample size of 160, there should be 80, 40, 40 in the three departments, giving the same proportions in the sample as in the population. The cost is $K\{(80 \times 0.81) + (40 \times 0.36) + (40 \times 0.64)\} = 104.80K$

(ii)

$Department$	Ni	Si	$\sqrt{Ci}$	$\frac{NiCi}{\sqrt{Ci}}$
$BusinessandManagement$	1200	2.7	0.9	3600
$ScienceandTechnology$	600	1.8	0.6	1800
$ArtandDesign$	600	1.6	0.8	1200

The sample sizes n_1, n_2, n_3 are proportional to $\frac{NiCi}{\sqrt{Ci}}$, and are therefore$\frac{3600}{3600}k = \frac{6k}{11};\ \frac{3k}{11};\ \frac{2k}{11}$. We determine k by considering total cost. This is $\frac{k}{11}[(6 \times 0.81) + (3 \times 0.36) + (2 \times 0.64)]$ which is $\frac{k}{11}(7.22)$, and cannot be greater than E100. Solving $\frac{k}{11} \times 7.22 = 100$ gives $k = 152035$ and sample sizes are therefore $n_1 = \frac{6k}{11} = 83.10;\ n_2 = 41.55;\ n_3 = 27.70$, so take $n_1 = 83,\ n_2 = 41,\ n_3 = 27$ to keep within budget. (n_3 could actually go up to 28-or n_2 to 42, but not both)(cost of $83 + 41 + 27$is 99.27k.)

6 (i) Questions can ne presented in pre-coded form (boxes to tick); an interviewer can code the answers as a respondent gives them; or answers can be given in full or the questionnaire and then coded in the office where analysis is done.

(ii) At stage 1, the respondent is not subject to interviewer bias, and answers need no interpretation by an interviewer; the data are ready immediately–But the respondents may not all fully understand the coding provided, and do not have any opportunity to give more detailed information.
At stage 2, the interviewer is present when answers are given and can interpret them, or clarify questions, in with training that has given–But can bias the respondent, or interpret complicated or detailed answers in live with personal prejudices, or the questionnaire can be completed too quickly with no chance to check it.
At stage 3, the coding should be done objectively, carefully without time constraint, can be checked, and additional information noted–But records may be incomplete, even lacking enough for accurate coding, any bias of respondent or interviewer cannot be detected.

(iii) Codes should be exhaustive and mutually exclusive (e.g.the income question in Question 4); and the question should be quite clear so that there is no doubt what information is required (e.g.whose income)

(iv) Tick the box which shows your highest educational attained:

1 □ *Postgraduate degree*
2 □ *First degree*
3 □ *Professional Qualification equivalent to degree − please specify*_________
4 □ *Higher National Diploma*
5 □ *Higher National Certificate*
6 □ *A − level*
7 □ *GNVQ*
8 □ *GCSE*
9 □ *Other − please specify*____________________

7 (i) The part of the table to be used may be chosen tossing a coin. H/T locates which side of page(left/right); H/T then for top/bottom; H/T again for moving up or down; H/T for reading left to right or the opposite;within the chosen section a starting point is found by locating row and column numbers,by an extension of these methods or by choosing two random digits from a RND pocket calculator output.

(ii) There are 488 pages, So triplets of random digits are required. To save wastage,number the pages 001-488 and make 501-988 correspond with there also: e.g.636 means page 136,072 means page72, etc. Every page is associated with two triplets; 000, 489-500, 989-999 are not used.

If a page chosen by this process is entirely tables or diagrams, a new choice is made.

For a line, make the digits correspond in pairs, 01-44 is the same as 51-94, and 00, 45-50, 95-99 are not used. If there are less than lines or a chosen page, this process may need to be repeated until an existing line is located.

Finally, take digits in pairs, discard 00 and 91-99, let 01-15 denote the words or a live(as many as there may be)and equate 16-30, 31-45, 46-60, 61-75, 76-90 with 01-15. Select a pair of digits and if this fails to locate a word or the chosen live then take a further pair. Alternatively, the whole selection process may be restarted if it breaks down at any stage. Using the right-hand column, reading down,963286095329631 gives 963 02 86,which is 463 line 2 word11 and 095 32 31, which is 95 line 32 word1(96 is discarded).

8 (i) In sampling households or smallholdings in all the villages in an agricultural area, there may not be an available frame of individuals. The first step is to make a complete list of villages. A random selection of villages is made, and a frame constructed for each selected villages.Then a ransom choice of households is made from each of villages. This is a clusters sample with two stages.A simple random sample would require a complete frame for the whole area, and even if this was available the time and cost of going round the whole area to find each selected individual could well be prohibitive. A cluster sample would work well if villages are similar to the other, but less well if variability between villages was high and a very untypical was chosen.

(ii) In rapid surveys of people's opinions, especially for marker research or political purposes, simple random samples may be too slow and some to carry out. Quota samples using well trained interviewers can give good results provided the quotas do reflect the population structure adequately and the interviewers do not deliberately avoid some groups of people when choosing the individuals to make up their quotas of each section of the population. If the population has not been divided into sections that are relevant to the purpose of the enquiry the results can be misleading.

Paper II

1 (i) A ratio is a number of the form X/Y, where X and Y are counts of the same type element, e.g. X is the number of passenger carried on a railway train during a rush-how and Y is the number carried on the same route at a quiet period; such as $X = no.$ of passengers or the 08.00 train and $Y = no.$ on the 11.00 train. Suppose $X = 200$ and $Y = 40$; then the ratio is 5 to 1,or 5.0 .

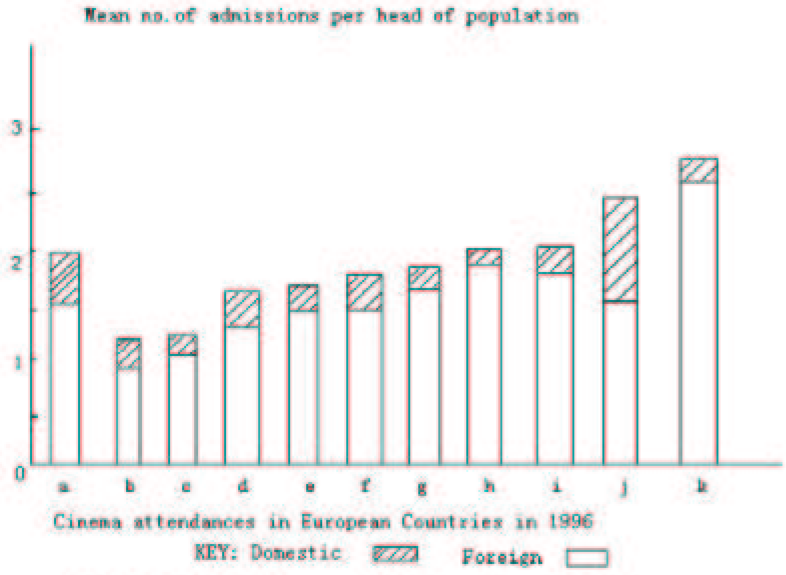
A rate is also of the form X/Y, but X is now include in Y; e.g. the rate of turn over of employees in a company, X is the number leaving during a year and Y is the number on the payroll at the beginning of the year. Suppose $X = 15$ and $Y = 345$; then the rate is $\frac{15}{345} = \frac{1}{23} = 0.0435$ or 435%or 1 in 23.

(ii)

	(a) Percentagesare :		*(b) and Means :*		
	Domestic	*Foreign*	*Domestic*	*Foreign*	*All*
Belgium	5.0	95.0	0.10	1.97	2.08
Denmark	17.0	83.0	0.32	1.56	1.88
Finland	10.6	89.4	0.11	0.97	1.08
France	37.4	62.6	0.88	1.48	2.36
Germany	15.1	84.9	0.25	1.38	1.63
Italy	24.5	75.5	0.39	1.21	1.60
Netherlands	5.7	94.3	0.06	1.03	1.09
Spain	8.9	91.1	0.24	2.42	2.66
Sweden	19.3	80.7	0.34	1.41	1.75
UK	$\frac{10.4}{18.7}$	$\frac{89.6}{81.3}$	$\frac{0.22}{0.36}$	$\frac{1.89}{1.58}$	$\frac{2.11}{1.94}$

2 A component bar chart is best because : (i) it shows clearly the relation sizes of the two components, (ii) it allows ordering in some convenient way such as by total size of the overall attendance, (iii) it shows actual sizes of the components,rather than expressing them as percentages of the whole.

3 Let X be mean no of admissions per head to domestic films and Y be average price of cinema



Cinema attendances in European Countries in 1996

a: All, b: Finland, c: Netherlands, d: Italy, e: Sweden, f: Denmark, g: Belgium, h: United Kingdom, i: France j:Spain

admission. These must be ranked, and the difference in ranks, d, is calculated.

Country	*B*	*D*	*Fi*	*Fr*	*G*	*I*	*N*	*Sp*	*Sw*	*UK*
Rank of X	2	7	3	10	6	9	1	5	8	4
Rank of Y	5	3	9	8	6	4	7	1	10	2
d	-3	4	-6	2	0	5	-6	4	-2	2

Spearman's $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6\times150}{10\times99} = 1 - 0.8091 = 0.091$

4 Spain shows the highest mean attendance per head of population; it also has the lowest average admission price. France has by far the highest attendance for domestic films, probably because the film industry is much larger than in other countries; this high attendance is in spite of having a fairly high admission price. The two countries, Sweden and Finland, which have the highest prices, seem to have different levels of attendance: Finland is the lowest of all, but Sweden is not far from average, Finland and Netherlands have some way the lowest mean admission number; the Netherlands price is the fourth highest. The correlations between mean number of admissions to domestic and foreign films,and between domestic numbers and price, are negligible; the first of these results is due to the very small numbers for domestic everywhere except France, so that the variation is nearly all in foreign admissions(for which France is fairly low!). The same reasons go for the record result, in spite of the comments made above.The correlation 3(ii)is largely due to France and Sweden. In fact the over riding influence is the high figure for "domestic"attendance in France; but otherwise the levels for most other countries are not very different.

5 There is one question of each type 'easy', 'moderate', 'different'.

$$P(correct|easy) = \frac{2}{3};\ P(correct|moderate) = \frac{1}{2};\ P(correct|different) = \frac{1}{5}.$$

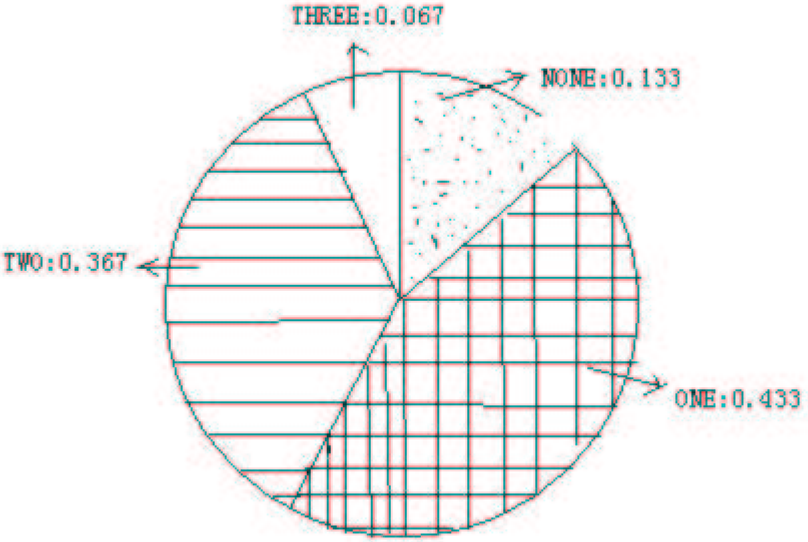
(i) $P(0) = P(all\ wrong) = (1-\frac{2}{3})(1-\frac{1}{2})(1-\frac{1}{5}) = \frac{1}{3}\times\frac{1}{2}\times\frac{4}{5} = \frac{2}{15} = 0.133$

$P(1) = (\frac{2}{3}\times\frac{1}{2}\times\frac{4}{5}) + (\frac{2}{3}\times\frac{1}{2}\times\frac{4}{5}) + (\frac{2}{3}\times\frac{1}{2}\times\frac{1}{5}) = \frac{4}{15} + \frac{2}{15} + \frac{1}{30} = \frac{13}{30} = 0.433$

$P(2) = \frac{2}{3}\times\frac{1}{2}\times\frac{4}{5} + \frac{2}{3}\times\frac{1}{2}\times\frac{1}{5} + \frac{1}{3}\times\frac{1}{2}\times\frac{1}{5} = \frac{4}{15} + \frac{1}{15} + \frac{1}{30} = \frac{11}{30} = 0.367$

$P(3) = \frac{2}{3}\times\frac{1}{2}\times\frac{1}{5} = \frac{1}{15} = 0.067\ \ Check\ Sum = 1$

(ii) The angles
0: 48^o; 1: 1
Probabiliti



6 (i) The weights w_i are the figures in "1995: no.of hours" and the costs x_i are those in the 1995 column.

$\sum w_i = 55130.\ \sum w_i x_i = (6835\times 117) + (7471\times 80) + \cdots + (8784\times 3.3) = 1657241.3$

The weighted average cost per hour is $\frac{1657241.3}{55130} = 30.061$

(ii) Laspeyres Indices are $100\times\sum w_{94}x_{95}/\sum w_{94}x_{94}$

For Television:

$$\frac{100\times(6843\times 117 + 6960\times 86)}{6843\times 112 + 6960\times 89} = \frac{139919100}{1385856} = 100.96$$

For Radio:

$$\frac{100 \times (8723 \times 2.7 + 8735 \times 3.6 + 6728 \times 7.4)}{8723 \times 2.6 + 8735 \times 3.8 + 6728 \times 7.1 + 7449 \times 10.2 + 8740 \times 3.1} = \frac{21184180}{206715.4} = 102.48$$

Radio had the greater relative increase.

7 (a) $n = 20$ $\sum x_i = 336.0$ $\sum x_i^2 = 6700.68$ $s^2 = \frac{1}{19}(6700.68 - \frac{336}{20})^2 = 55.5726$ $\overline{x} = \frac{336}{20} = 16.8$ $so\ s = 7.455$ $\%CV = \frac{100s}{\overline{x}} = 44.37\%$

(b) Inter-subject variation is the variation in the same measurement taken on several different units, such as in (a) above. It indicates how variable the measurement (of body fat, in the example) is over a sample of people from a defined population-here man aged 22-35. Intra-subject variation could be measured if the same measurement (body fat) was taken more than once, under the same conditions, as each unit(person). This cannot be used for comparisons between people; it measures how reproducible the measurement is, or how much it varies according to either variation in the measurement technique or natural short-term variation in the person. In the case of body fat it should usually reflect technique only.
The two types of variation is normally much greater than intra-subject variation. In future surveys there may be no be no need to measure intra-subject variation.

8 (i) Different Data-Trend(cm)

$Year$	$Jan - Feb$	$Mar - Api$	$May - June$	$July - Aug$	$Sept - Oct$	$Nov - Dec$
1979	−1542	−912	2431	1340	−37	−1120
1980	−1725	−1298	3302	1440	−670	−1317
1981	−1785	−1529	2122	3690	−814	−1469
average	−1684.0	−1246.3	2618.3	2156.7	−507.0	−1302.0

$Sum = 35.7$ $Correction = -\frac{35.7}{6} = -5.95$ standardized seasonal components are: $-1689.9\ -1252.3\ 2612.4\ 2150.7\ -512.9\ -1307.9$.

(ii) Residual component=observed value-trend-seasonal component.
1981 -95.1 -276.7 -490.4 1539.3 -301.1 -161.1
Residual numerically greater than 400