THE ROYAL STATISTICAL SOCIETY

ORDINARY CERTIFICATE
(2 papers)

SOLUTIONS 1997

Note:

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

The Royal Statistical Society, 12 Errol Street, London EC1Y 8LX, UK

Paper I

1. (i) A report should contain a clear statement of the aims of the survey. Details of pre-testing / pilot trial should be included.

   (ii) The target population should be specified clearly, such as whether all age-groups were included, male and female, head of household or all individuals.

   (iii) The details of time, dates, places covered in data collection, stratification into groups if any, should be given.

   (iv) The basis of the sampling frame should be explained, along with the sampling method used from it - random, systematic, stratified, multistage, cluster - were interviewers / enumerators used?

   (v) Any necessary comments about non-response, such as whether certain groups of people did not respond or whether certain questions or requires for information caused offence, should be made, with an indication of possible effects of this.

   (vi) List of contents, tables and graphs.

   (vii) Careful definition of technical terms used - e.g. "form", "smallholding", "income", "consumption", "employment".

   (viii) Tables and graphs where relevant and useful to the discussion, presented in the same form as previously where quarterly yearly, etc. Comparisons are to be made. Good, clear labelling necessary.

   (ix) Discussion should be arranged in sections, with headings and subheadings as appropriate. Supplementary information, agricultural, economic, political, climatic, should be mentioned at the appropriate points.

   (x) Copies of the questionnaire, interviewers' instructions, rules for coding answers etc. should be given in an appendix, along with any large quantities of raw data that are to be published.

   (xi) Inferences, conclusions and projections should appear last in the report and be justified by cross-references to earlier sections.

   [Experience with particular types of survey may provide further ideas.]

2. (a) Systematic sampling from a list is simpler and much quicker than random sampling because only the starting point has to be determined at random; hence it is much cheaper. However, if there are any regular or cyclic trends through the list that is being used, it is possible for a systematic sample to get "in phase" with these. This can cause bias. Also if the sample size is not an exact fraction of the population size some units may not have exactly the same probability of selection as others.

(b) There are $N = 24$ pupils, and a sample of $n = 8$ is required; this is 1/3 of the class. One of A, B, C must be selected at random, e.g. by throwing a six-sided die and choosing A when the score is 1 or 2, B for 3 or 4, C for 5 or 6. Then every third name in the list is taken; e.g. B, E, H, K, N, Q, T, W.

3. (a) (i)(ii) In stead of selecting individual sample members directly, they may be grouped into sets, e.g. all the dwellings in one street or all the farms in one village or all the schools in one country, and a random selection of streets or villages or countries made first, followed by selection of individuals from these streets etc. This is two-stage sampling, and more stages may be included in the selection to give multi-stage sampling schemes, e.g. cities or large towns in a country at stage 1, areas or wards in the chosen city at stage 2, and streets at stage 3 - we could then end with individual houses or, households at stage 4. A sampling frame must be available in detail for all the units in stage 4, but is not always essential in full at the earlier stages provided the existence of all units at that stage is known.

(b) (i) Using as a two-stage example the sampling of a region by villages from a list of all villages, followed by farms within each selected village, villages may be classified into groups for, e.g., climate, communications, accessibility or type of agricultural activity (crops, animals etc.) and simple random samples may be taken from each group. In the selected villages, farms could be classified "large" or "small", and simple random samples from each of these "strata" could be taken.

(ii) In cluster sampling, all the farms in each selected village would be used.

(c) Cluster sampling does not require a frame for each village, since all the farms are to be used; only care is needed to locate them all. Hence the cost of making a frame is eliminated. On the other hand, when strata are relatively homogeneous within themselves, cluster sampling may be wasteful of units and also introduce bias or increase sampling errors overall.

4. S_i may be estimated from the pilot survey as $\{\frac{1}{n-1}(\sum x_i^2 - \frac{(\sum x_i)^2}{n})\}^{1/2}$.

These estimates are: $S_D = \sqrt{\frac{1}{99}(14107.25 - \frac{1161.4^2}{100})} = 2.50$;

$S_C = \sqrt{\frac{1}{99}(16705.62 - \frac{1229.7^2}{100})} = 4.00$; $S_I = \sqrt{\frac{1}{99}(33058.44 - \frac{1634.7^2}{100})} = 8.00$.

If $n_i = KN_iS_i$, we have
$$n_D = K(800000)(2.5) = 2K \times 10^6$$
$$n_C = K(150000)(4.0) = 6K \times 10^5$$
$$n_I = K(50000)(8.0) = 4K \times 10^5,$$

i.e. $n_D : n_C : n_I = 20 : 6 : 4 = 10 : 3 : 2 = \frac{10}{15} : \frac{3}{15} : \frac{2}{15}$, so that the actual sample sizes are 10,000, 3,000, 2,000.

5. (a) In a cross-sectional survey, data are collected from each unit once only, such as the yield from a grain crop at harvest time. A longitudinal survey collects data from the same units repeatedly, e.g. weekly expenditure patterns for households taking part in a consumer survey.

   (b) Longitudinal surveys give better estimates of trends by reducing personal variations. Subjects serve as their own "controls" to indicate the effects of any particular events during the survey, e.g. price changes of competing products. However, some individuals may drop out of a longitudinal survey and bias may result from having incomplete data. Also the familiarity gained by regular participation way eventually make people "non-representative" of the general population.

6. Outline contents for the sections:

   (i) Select a random sample from the available database for telephone sampling - must do this in the same way as selecting the sample from a frame such as an electors' list for interviewing. But the database will be different - possibly only a telephone directory - unless a previous 'consumer survey' database has been provided by an agency or gathered by the research group in its previous work. If only a telephone directory, 'ex-directory' people are excluded, which could cause bias. To cover an area by stratified sampling it would be less easy to use another database than to begin with an electoral list. If a selected sample member is not available, telephone sampling avoids the waste of visiting and getting no reply. In both methods, the sample originally chosen should be used even if repeated attempts to contact are necessary, with 'reserve' replacements only made if a member is genuinely not now in the population. This may be difficult to find out by telephone sampling. Basic methods are thus similar but telephone sampling may be done on a less representative database, without this becoming obvious during a survey.

   (ii) Telephone responses may be more abrupt - immediate refusal without opportunity for explaining the survey may be more likely, especially if the same people are surveyed too often. Face-to-face interviews by trained interviewers will extract more accurate information if a proper friendly relationship can be established and the interviewer senses whether a question has been properly understood. Some groups of people, e.g. the elderly, or those who are busy when the telephone rings, may give inaccurate, hasty or ill-considered answers. With interviews, especially when arranged in advance, there will be less danger of being given deliberately wrong or valueless information, simply in order to get rid of the caller.

(iii) Refusal / non-response may be higher by telephone, but it is also possible to make more attempts to contact this way, so there may be some balance. Costs of interviews are much higher, so there can usually be fewer of them in a survey than in a telephone method. Biases may be less easy to detect by telephone, as mentioned above; also the database may be less complete. If some detail is asked for, which the respondent cannot remember but has to look up, face-to-face interviews give more opportunity for this to be done. Even if an interviewer completes the form, the face-to-face respondent can have a copy if this will help to clarify questions and alternative answers. By telephone, it is not possible to give many answers to choose from, as most of the list will be forgotten.

7. (a) This is a non-random error due to causes other than the sampling process, selection etc., which has been used.

(b) Poor measuring equipment, poor questionnaire design, poor interviewing technique, poor coding of possible answers, poor data - entry procedures, lack of clarity / understanding of questions.

Poor measuring equipment, e.g. when weighting agricultural crop samples, adds a 'zero error' to measurements taken. Poor questionnaire design leads to questions being answered wrongly because they are misunderstood. Poor interviewing technique can lead to wrong answers either because the interviewer has put the question in a confusing way or has biased it towards a particular answer. If coding omits possible answers, respondents will either tick the wrong box or not answer at all. Poor data-entry introduces unnecessary errors into the data used for processing.

A pilot survey and proper training of interviewers can remove several of these sources of bias, and data-entry checking should avoid loss of quality at that stage. Calibration of measuring equipment should be a routine part of setting up a survey. Attention to previous similar surveys could point to likely sources of error, and serious attempts could be made to avoid these.

8. Optical character recognition scanners, optical mark readers and graphics or digital scanners can be used. Data recorded on sheets may be scanned, data measured by weighing instruments may be automatically read into a database either directly or by scanning the recording display screen, characteristics of crops may be recorded. Time-saving and accuracy can be advantages of these methods. Disadvantages can be unclear recording sheets which do not scan well, extraneous substances which may cause an impression, especially in field work, visual characteristics way need special methods of calibration

to establish a scale of measurement.

Paper II

1. (a) (i) 54700. (ii) 0.0347. (iii) 600 000.

   (b) (i) 1325+670+700=2695 but can only be expressed to the level of accuracy of the least accurate figure (700), so is 2700.

   (ii) $1.88 \times 4.3 = 8.084$, but the number of significant figures should not exceed the minimum number in the figures in the product, which is 2, so quote 8.1.

2.

Width-upper limit (mm)	Cumulative frequency
5.0	15
10.0	40
14.0	62
25.0	86
50.0	98
75.0	100

(i) Graph - see next page.

(ii) Median at 50^M observation which is 11.0 mm. (read to nearest 0.5). 90^M percentile is 29.5 mm. 10^M percentile is 3.0 mm.

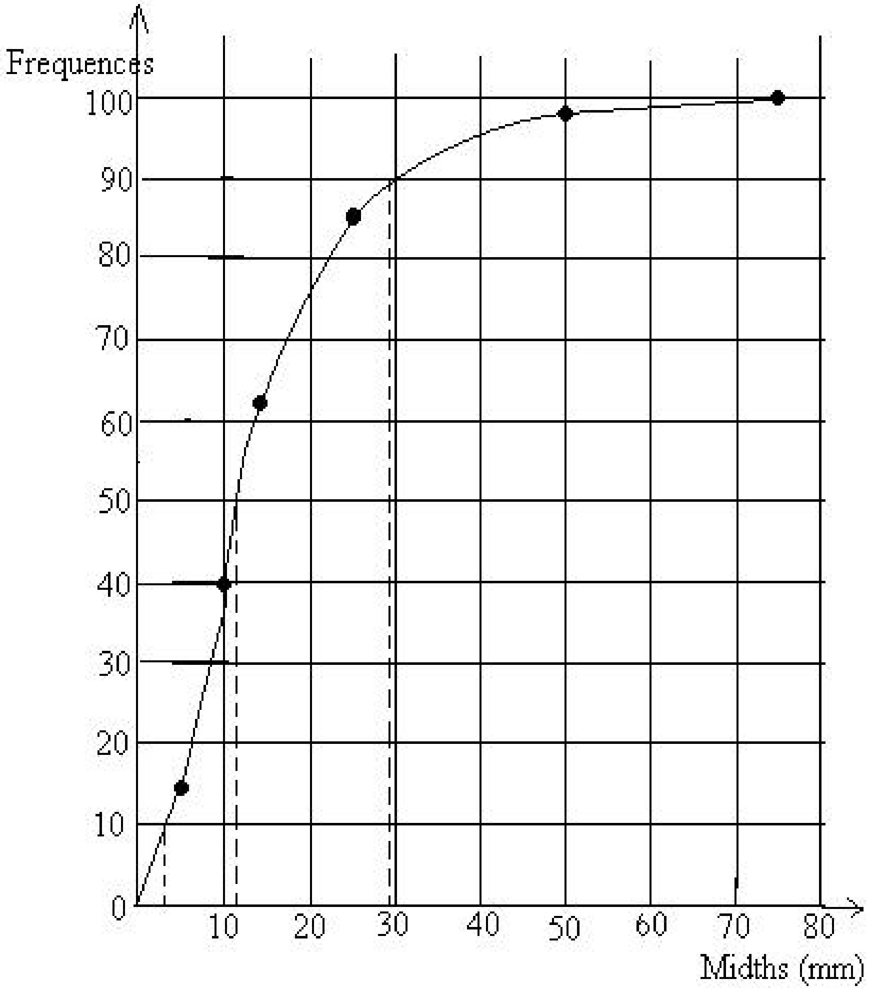
3.

Width-midpoint of interval(mm)	Frequency of in interval		
X	f	f_x	f_x^2
2.5	15	37.5	93.75
7.5	25	187.5	1406.25
12.0	22	264.0	3168.00
19.5	24	468.0	9126.00
37.5	12	450.0	16875.00
62.5	2	125.0	7812.50
	100	1532.0	38481.50

Mean= $\frac{\sum f_x}{\sum f} = 15.32$mm. Variance $= \frac{1}{99}(\sum f_x^2 - \frac{\{\sum f_x\}^2}{\sum f})$,

hence Standard Deviation $= \sqrt{\frac{1}{99}(38481.5 - \frac{1532^2}{100})} = 12.31$mm.

4. The table should be laid out at right angles to the present plan, so that all the


Midths of a Random Sample of 100 books: Cumulative Frequances

country headings can go in one line:

Belgium ··· ··· ··· ··· ··· ··· UK
Population
⋮
% ··· Industry

Visual comparison across countries is now easy.

These may not be need for all the horizontal ruled lines, nor perhaps the vertical ones, so long as the levels and margins are kept in line, as now; extra neatness may help visual comparisons.

Row labelling need not be so detailed: abbreviations like sq. km., OOO's and footnotes for (per 1000 population) will improve impact and readability.

Instead of arranging countries alphabetically, it may be useful to order them according to a particularly important variable, such as population. An addition to this could be to group them in any relevant way, depending on the purpose for which the table is used; leave an extra space between groups.

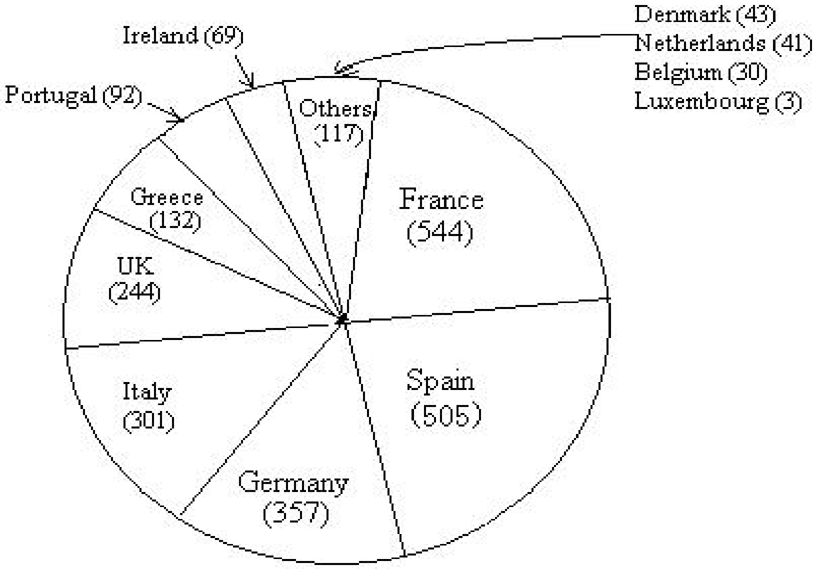
Population figures could be expressed in millions rather than thousands: this would make them easier to read (and would not lose important information). The same might be done with employment totals.

A possible case could be make out for having rows and columns the other way round:

	Population	···	···	···	···	%Employment in Industry
Belgium						
⋮						
UK						

There are arguments both ways on the value of this, depending on the aim of the report. Countries, or groups of countries, may be easier to compare this way.

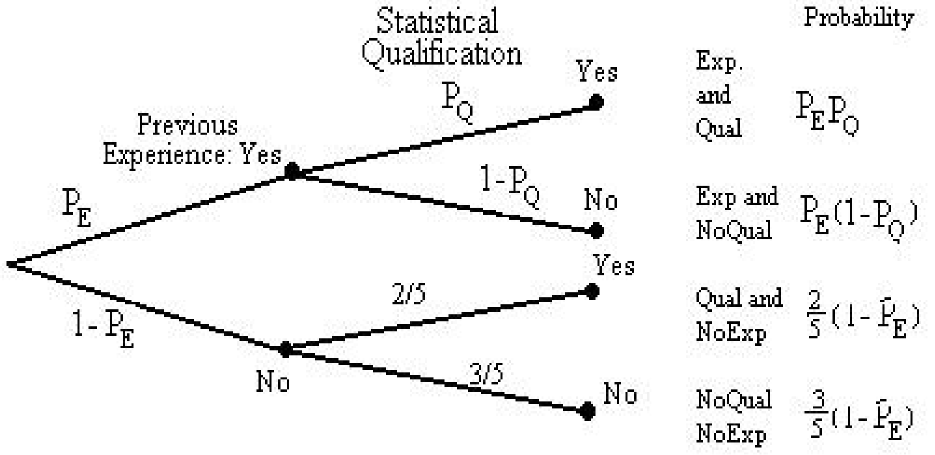
5.



Area of the European Community, 1991 ('OOOs aq. km.)
Source: SOEC(1993).

	Area(sq.km.)	%of whole	Angle(^{0})
France	543965	23.0	82.9
Spain	504790	21.4	77.0
Germany	356854	15.1	54.4
Italy	301287	12.8	45.9
UK	244111	10.3	37.2
Greece	131957	5.6	20.1
Portugal	91971	3.9	14.0
Ireland	68895	2.9	10.5
Other	117191	5.0	17.9
	2361021	100.0	359.9

6. (i)



(ii) (a) Given that $\frac{3}{5}(1-P_E)=\frac{1}{3}$, we have $1-P_E=\frac{5}{9}$ or $P_E=4/9$.

(b) Given that $P_EP_Q+\frac{2}{5}(1-P_E)=\frac{1}{2}$, we have $\frac{4}{9}P_Q+\frac{2}{5}\cdot\frac{5}{9}=\frac{1}{2}$, i.e. $\frac{4}{9}P_Q=\frac{1}{2}-\frac{2}{9}=\frac{5}{18}$ or $P_Q=\frac{9}{4}\times\frac{5}{18}=5/8$.

(c) If the number of applications was N, then $NP_E(1-P_Q)=6$, i.e. $N\cdot\frac{4}{9}\cdot\frac{3}{8}=6$ or $N=\frac{6\times8\times9}{4\times3}=36$.

7. $n=25$ pairs of records. $\bar{S}=\frac{616.23}{25}=24.65$. $\bar{J}=\frac{153.81}{25}=6.15$. For the whole population of 25 competitors, the variance may be calculated using n as divisor (instead of $(n-1)$ if it had been a sample).

$$\sigma_S^2=\frac{1}{25}(15212.14-\frac{616.23^2}{25})=0.9025,\quad \sigma_S=0.950,$$
$$\sigma_J^2=\frac{1}{25}(951.70-\frac{153.81^2}{25})=0.2160,\quad \sigma_J=0.465.$$

(i) Coefficients of variation are, for S: $\frac{100\times0.95}{24.65} = 3.85\%$; for J: $\frac{100\times0.465}{6.15} = 7.56\%$; The value for Long Jump is greater.

(ii) See graph on next page.

(iii)

$$\begin{aligned} r = \frac{\frac{1}{n}\{\sum S_iJ_i - (\sum S_i)(\sum J_i)/n\}}{\sqrt{\sigma_S^2\sigma_J^2}} &= \frac{1}{25}\cdot\frac{1}{0.95}\cdot\frac{1}{0.465}(3782.28 - \frac{616.23\times153.81}{25}) \\ &= -0.816. \end{aligned}$$

(iv) As the length of jump made by a competitor decreases, so the corresponding length of time taken for the sprint increases. Competitors who are better performers for one tend also to be better for the other, making longer jumps and having shorter sprint times. The high (negative) correlation coefficient shows this, and the graph further shows that one competitor is, in some distance, best for both, and at the other extreme there are also two that are away from the main group.

8.

Ingredient	Price(b)		Weight ×	Price		
	1991(P_1)	1992(P_2)	WP_1	WP_2	P_2/P_1	WP_2/P_1
Juice	84.7	82.4	423.5	412.0	0.973	4.865
Gereal	223.5	239.5	5587.5	5987.5	1.072	26.800
Milk	55.1	56.3	275.5	281.5	1.022	5.110
Eggs	9.2	9.3	138.0	139.5	1.011	15.165
Bread	89.1	89.2	891.0	892.0	1.001	10.010
Butter	239.9	246.8	1199.5	1234.0	1.029	5.145
Marmalade	154.5	162.0	2317.5	2430.0	1.049	15.735
Tea	514.1	501.0	10282.0	10020.0	0.975	19.490

Seaview uses the weighted aggregative index

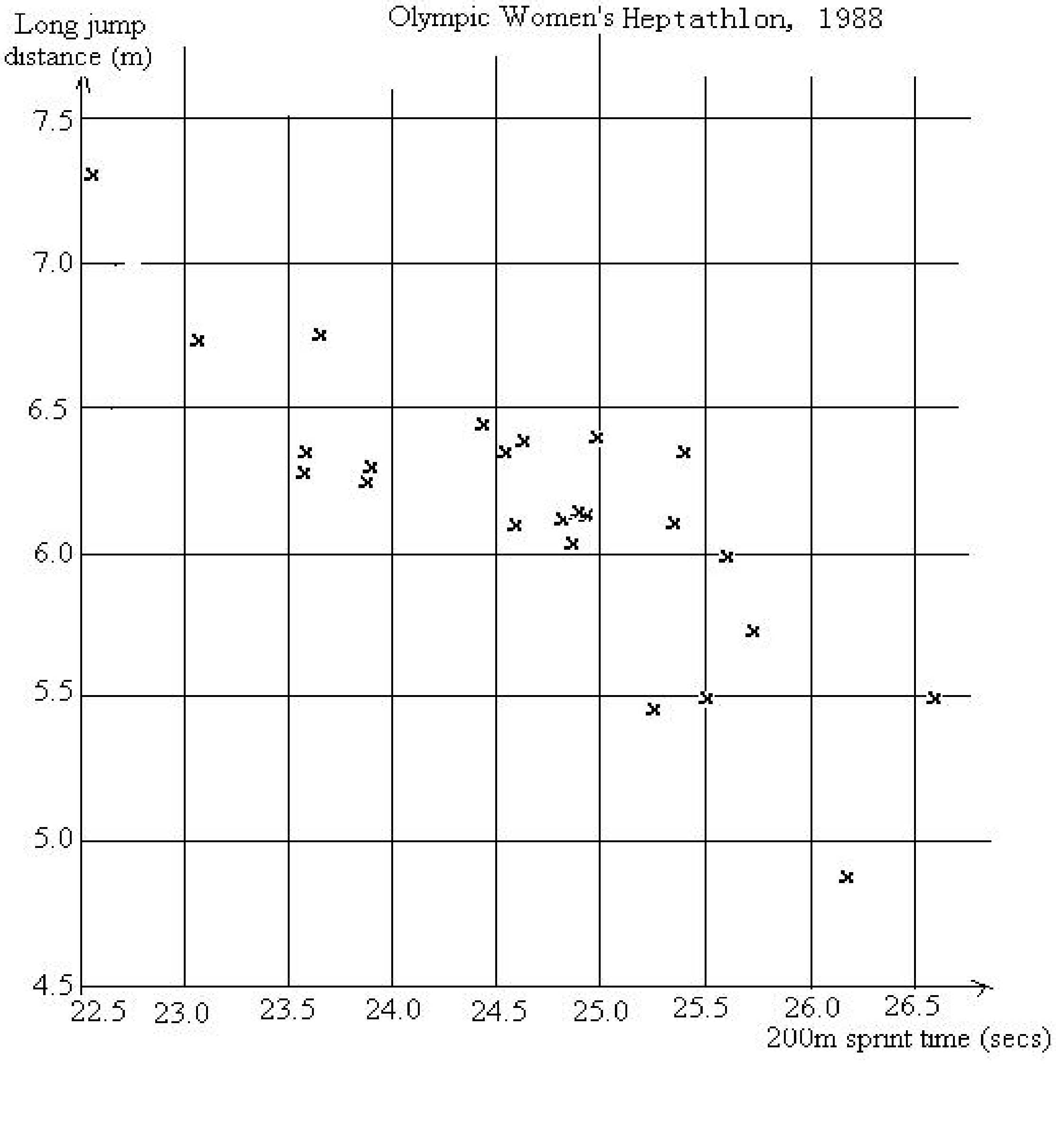
$$100\frac{(\sum WP_2)}{(\sum WP_1)} = 100\cdot\frac{21396.5}{21114.5} = 101.3.$$

Grand uses the weighted average of price relatives

$$100\frac{(\sum WP_2/P_1)}{\sum W} = 100\cdot\frac{102.320}{100} = 102.3.$$

Grand was more expensive.

Prices would be £3.55 (Seaview) and £3.58, which might be rounded to the nearest 5p at 3.60 (Grand).


Olympic Women's Heptathlon, 1988
Long jump
distance (m)
7.5
7.0
6.5
6.0
5.5
5.0
4.5
22.5
23.0
23.5
24.0
24.5
25.0
25.5
26.0
26.5
200m sprint time (secs)